Spatial adaptation in heteroscedastic regression: Propagation approach

Nora Serdyukova

May 13, 2011

Abstract

The paper concerns the problem of pointwise adaptive estimation in regression when the noise is heteroscedastic and incorrectly known. We use the method of local approximation including as a particular case the local polynomial smoothing. Specifically, the model with unknown mean and variance is approximated by a local linear model with an incorrectly specified covariance matrix.

Adaptive choice of degree of localization in this case can be understood as a choice of an appropriate parametric model from a given collection. For the selection from the family of models we employ based on Lepski's method the FLL technique recently suggested in Katkovnik and Spokoiny (2008). The problem of the choice of certain parameters in this type of procedures was addressed in Spokoiny and Vial (2009). The authors called their approach to the calibration of the parameters "propagation". We developed and justified the methodology for the heteroscedastic case in the presence of noise misspecification. The analysis shows that the adaptive procedure allows a misspecification of the covariance matrix with a relative error of order $(\log n)^{-1}$, where n is the sample size.

1 Introduction

Consider a regression model

$$\boldsymbol{Y} = \boldsymbol{f} + \Sigma_0^{1/2}\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n) \tag{1.1}$$

with response vector $\boldsymbol{Y} \in \mathbb{R}^n$ and an unknown covariance matrix $\Sigma_0 = \operatorname{diag}(\sigma_{0,1}^2, \ldots, \sigma_{0,n}^2)$. This model can be written as

$$Y_i = f(X_i) + \sigma_{0,i}\,\varepsilon_i, \quad i = 1, \ldots, n \tag{1.2}$$

with design points $X_i \in \mathcal{X} \subset \mathbb{R}^d$. Given a point $x \in \mathcal{X}$, the target of estimation is the value of the regression function $f(x)$ at the point x.

The idea is to replace model (1.2) by a local parametric model

$$y_i = f_{\boldsymbol{\theta}}(X_i) + \sigma_i\,\epsilon_i, \quad i : X_i \in U_h(x), \tag{1.3}$$

where $U_h(x) \stackrel{\text{def}}{=} \{t : \|t - x\| \le h\}$ and $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ is an unknown finite-dimensional parameter. Then employing one of the well-developed parametric methods we can estimate $\boldsymbol{\theta}$ by $\widetilde{\boldsymbol{\theta}}(y_1, \ldots, y_d; x)$, and then use the estimator $\widetilde{f}_{\widetilde{\boldsymbol{\theta}}(Y_1,\ldots,Y_d)}(x)$ based on the observations from the

"true" model (1.2) for estimation of $f(x)$. Therefore we have to choose the local model (the collection of estimators $\{f_{\boldsymbol{\theta}}(\cdot), \boldsymbol{\theta} \in \Theta\}$) and the appropriate degree of locality h. This method of local approximation originated from [14], [15], [16], [17]. In what follows we shall consider approximation by local linear models of the following type:

$$y_i = \Psi_i \boldsymbol{\theta} + \sigma_i \, \epsilon_i, \quad i : X_i \in U_h(x), \tag{1.4}$$

where $\Psi_i = \Psi(X_i) = (\psi_1(X_i - x), \ldots, \psi_p(X_i - x))^\top$ is a vector of basis functions $\{\psi_j(\cdot)\}$ which already are fixed. The main issue then is the choice of the appropriate bandwidth h such that the estimator

$$\widetilde{f}_{\widetilde{\boldsymbol{\theta}}_h}(x) \stackrel{\text{def}}{=} \sum_{j=1}^{p} \widetilde{\theta}_h^{(j)}(x) \psi_j(0) \tag{1.5}$$

built on the base of the localized data would be a relevant estimator for $f(x)$. For this purposes the bandwidths selection should be done in a data-driven way, and the adaptive selection from the family $\{\widetilde{f}_{\widetilde{\boldsymbol{\theta}}_h}(\cdot)\}_{h>0}$ for fixed basis is equivalent to the adaptive choice of bandwidth. Notice also that the coefficients $\theta^{(j)}(x)$ as well as their estimators depends on x and should be calculated for every particular point of interest x. On the other side the localization reduces influence of the choice of the functions $\{\psi_j(\cdot)\}$ allowing to use simple collections.

Moreover, in our set-up the covariance matrix Σ_0 is not assumed to be known exactly, and the approximate model used instead of the true one reads as follows:

$$\boldsymbol{y} = \boldsymbol{\Psi}^\top \boldsymbol{\theta} + \Sigma^{1/2} \boldsymbol{\varepsilon}, \tag{1.6}$$

where $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_n)$ is a $p \times n$ design matrix and $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, $\min\{\sigma_i^2\} > 0$ is available to a statistician. Thus the model is misspecified in two places: in the form of the regression function and in the error distribution.

The proposed approach includes the important class of polynomial regressions, see [6], [17], [24], [28]. For example in the univariate case $x \in I\!R$, due to the Taylor theorem, the approximation of the unknown function $f(t)$ for t close to x can be written in the following form: $f_{\boldsymbol{\theta}}(t) = \theta^{(0)} + \theta^{(1)}(t-x) + \cdots + \theta^{(p-1)}(t-x)^{p-1}/(p-1)!$ with the parameter $\boldsymbol{\theta} = (\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(p-1)})^\top$ corresponding to the values of $f(\cdot)$ and its derivatives at the point x, if they exist. The matrix $\boldsymbol{\Psi}$ then consists of the columns $\Psi_i = \bigl(1, X_i - x, \ldots, (X_i - x)^{p-1}/(p-1)!\bigr)^\top$ and corresponds to the well known polynomial smoothing. If the regression function is sufficiently smooth, then for any t close to x, up to a reminder term $f(t) \approx f_{\boldsymbol{\theta}}(t)$, and the estimate of $f(x)$ at the point x is given by $\widetilde{f}(x) = f_{\widetilde{\boldsymbol{\theta}}(x)}(x) = \widetilde{\theta}^{(0)}$. See for more details on local polynomial estimatin [6] or [28]. The local constant fit at a given point $x \in I\!R$ is covered as well. In this case the design matrix $\boldsymbol{\Psi} = (1, \ldots, 1)$, and $f_{\boldsymbol{\theta}}(X_i) = \boldsymbol{\Psi}_i^\top \boldsymbol{\theta} = \theta^{(0)} = f_{\boldsymbol{\theta}}(x)$, $i = 1, \ldots, n$. This type of approximation in our set-up with known constant noise is treated in [18] and [27].

Nonparametric estimation in heteroscedastic regression under the L_2 losses was studied in [12], [13] and series of papers [8], [9], [10]. For estimation of the mean with L_2-risk in Gaussian homoscedastic model with unknown variance the penalties allowing to deal with the complexity of such a collection of models were proposed in [2]. However the problem of "local model selection" addressed in the present paper is quite different to the model selection in the sense of [3] and [25] related to estimation with global risk. The minimax pointwise estimation in heteroscedastic regression is in focus of [4].

2 Estimation procedure

2.1 Local parametric estimation

We shall perform the adaptive selection from a collection of K estimators corresponding to model (1.4) with different sizes h of neighborhood $U_h(x)$. Fix a point $x \in \mathbb{R}^d$ as a center of localization and a basis $\{\psi_j\}$. Let the localizing operator be identified with the corresponding matrix. For the next nonparametric step we will need a sequence of nested windows. Thus for every x the sequence of localizing schemes (scales) $\mathcal{W}_k(x)$, $k = 1, \ldots, K$ is given by the matrices $\mathcal{W}_k(x) = \mathrm{diag}(w_{k,1}(x), \ldots, w_{k,n}(x))$, where the weights $w_{k,i}(x) \in [0,1]$ can be understood, for instance, as smoothing kernels $w_{k,i}(x) = W((X_i - x)h_k^{-1})$. Let a particular localizing function $w_{(\cdot,\cdot)}(x)$ be fixed; the aim is to choose on the base of available data the index k of the optimal bandwidth h_k. To simplify the notation we sometimes suppress the dependence on the reference point x. Denote by

$$\mathbf{W}_k \stackrel{\mathrm{def}}{=} \Sigma^{-1/2}\mathcal{W}_k\Sigma^{-1/2} = \mathrm{diag}\left(\frac{w_{k,1}}{\sigma_1^2}, \ldots, \frac{w_{k,n}}{\sigma_n^2}\right), \quad k = 1, \ldots, K. \tag{2.1}$$

Let Θ be a compact subset of $\mathbb{R}^p$. Inside of any "window" given by $\mathbf{W}_k$, $k = 1, \ldots K$ we calculate the quasi-maximum likelihood estimator (QMLE) $\widetilde{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_k(x) = (\widetilde{\theta}_k^{(0)}(x), \ldots, \widetilde{\theta}_k^{(p-1)}(x))^\top$ of $\boldsymbol{\theta}$ defined as

$$\widetilde{\boldsymbol{\theta}}_k \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}), \tag{2.2}$$

where

$$\begin{aligned} \mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}) &= -\frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{\Psi}^\top\boldsymbol{\theta}\right)^\top \mathbf{W}_k \left(\boldsymbol{Y} - \boldsymbol{\Psi}^\top\boldsymbol{\theta}\right) + R \\ &= -\frac{1}{2}\sum_{i=1}^n |Y_i - \Psi_i^\top\boldsymbol{\theta}|^2 \frac{w_{k,i}}{\sigma_i^2} + R; \end{aligned} \tag{2.3}$$

R stands for the terms not depending on $\boldsymbol{\theta}$, and

$$\Psi_i = \Psi(X_i - x) = (\psi_1(X_i - x), \ldots, \psi_p(X_i - x))^\top.$$

If the $p \times p$ matrix $\mathbf{B}_k = \mathbf{B}_k(x)$ given by

$$\mathbf{B}_k \stackrel{\mathrm{def}}{=} \boldsymbol{\Psi}\mathbf{W}_k\boldsymbol{\Psi}^\top = \sum_{i=1}^n \Psi_i\Psi_i^\top \frac{w_{k,i}}{\sigma_i^2} \tag{2.4}$$

is positive definite ($\mathbf{B}_k \succ 0$), then

$$\widetilde{\boldsymbol{\theta}}_k = \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\boldsymbol{Y} = \mathbf{B}_k^{-1}\sum_{i=1}^n \Psi_i Y_i \frac{w_{k,i}}{\sigma_i^2}. \tag{2.5}$$

Recall that in the case of the polynomial basis the estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ is a local polynomial estimator of $\boldsymbol{\theta}(x)$ corresponding to the kth scale. In what follows we assume that $n > p$, and that $\det \mathbf{B}_k > 0$ for any $k = 1, \ldots, K$. Because $p = \mathrm{rank}(\mathbf{B}_k) \le \min\{p, \mathrm{rank}(\mathcal{W}_k(x))\}$ this requires the following conditions on the design matrix $\boldsymbol{\Psi}$ and the minimal localizing scheme $\mathcal{W}_1(x)$:

(A1) *The* $p \times n$ *design matrix* $\boldsymbol{\Psi}$ *is supposed to have full row rank, i.e.,*

$$\dim \mathcal{C}(\boldsymbol{\Psi}^\top) = \dim \mathcal{C}(\boldsymbol{\Psi}^\top \boldsymbol{\Psi}) = p.$$

(A2) *The smallest localizing scheme* $\mathcal{W}_1(x)$ *is chosen to contain at least* p *design points such that* $w_{1,i}(x) > 0$*, i.e.,* $p \le \#\{i : w_{1,i}(x) > 0\}$*.*

Assumption $(A2)$ in practise is automatically fulfilled, since, for example, in $\mathbb{R}^1$ it means that for the local constant fit we need at least one observation and so on. Usually it is intrinsically assumed that, starting from the smallest window, at every step of the procedure every new window contains at least p new design points.

The formulas (2.2) give a sequence of estimators $\{\widetilde{\boldsymbol{\theta}}_k(x)\}_{k=1}^K$. It was noticed in [1] that in the case when the true data distribution is unknown the QMLE is a natural estimator for the parameter maximizing the expected log-likelihood. That is, for every $k = 1, \ldots, K$, the estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ can be considered as an estimator of

$$\begin{aligned} \boldsymbol{\theta}_k^*(x) &\stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\, \mathrm{L}\left(\mathbf{W}_k, \boldsymbol{\theta}\right) && (2.6) \\ &= \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} (\boldsymbol{f} - \boldsymbol{\Psi}^\top \boldsymbol{\theta})^\top \mathbf{W}_k (\boldsymbol{f} - \boldsymbol{\Psi}^\top \boldsymbol{\theta}) \\ &= \mathbf{B}_k^{-1} \boldsymbol{\Psi} \mathbf{W}_k \boldsymbol{f} = \mathbf{B}_k^{-1} \sum_{i=1}^n \Psi_i f(X_i) \frac{w_{k,i}}{\sigma_i^2}. && (2.7) \end{aligned}$$

Recall that we do not assume $\boldsymbol{f} = \boldsymbol{\Psi}^\top \boldsymbol{\theta}$ even locally. It is known from [29] that in the presence of a model misspecification for every k the QMLE $\widetilde{\boldsymbol{\theta}}_k$ is a strongly consistent estimator for $\boldsymbol{\theta}_k^*(x)$, which also is the minimizer of the localized Kullback-Leibler [19] information criterion:

$$\begin{aligned} \boldsymbol{\theta}_k^*(x) &= \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n \mathbb{KL}\left(\mathcal{N}\left(f(X_i), \sigma_i\right), \mathcal{N}\left(\Psi_i^\top \boldsymbol{\theta}, \sigma_i\right)\right) w_{k,i}(x) \\ &= \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n |f(X_i) - \Psi_i^\top \boldsymbol{\theta}|^2 \frac{w_{k,i}(x)}{\sigma_i^2} \end{aligned}$$

with $\mathbb{KL}(P, P_{\boldsymbol{\theta}}) \stackrel{\text{def}}{=} \mathbb{E}_P\left[\log\left(\mathrm{d}P / \mathrm{d}P_{\boldsymbol{\theta}}\right)\right]$. For the properties of the Kullback-Leibler divergence see, for example, [28].

It follows from the above definition of $\boldsymbol{\theta}_k^*(x)$ and from (2.2) that the QMLE $\widetilde{\boldsymbol{\theta}}_k$ admits a decomposition into deterministic and stochastic parts:

$$\widetilde{\boldsymbol{\theta}}_k = \mathbf{B}_k^{-1} \boldsymbol{\Psi} \mathbf{W}_k (\boldsymbol{f} + \Sigma_0^{1/2} \boldsymbol{\varepsilon}) = \boldsymbol{\theta}_k^* + \mathbf{B}_k^{-1} \boldsymbol{\Psi} \mathbf{W}_k \Sigma_0^{1/2} \boldsymbol{\varepsilon} \quad (2.8)$$

$$\mathbb{E}\widetilde{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_k^*, \quad (2.9)$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n)$. Notice that if $\boldsymbol{f} \equiv \boldsymbol{\Psi}^\top \boldsymbol{\theta}$, then $\boldsymbol{\theta}_k^* \equiv \boldsymbol{\theta}$ for any k, and the classical parametric set-up takes place.

2.2 Adaptive bandwidth selection

Let a point $x \in \mathcal{X} \subset \mathbb{R}^n$, a basis $\{\psi_j\}$ and the method of localization $w_{(\cdot,\cdot)}(x)$ be fixed. The crucial assumption for the procedure under consideration to work is that the localizing schemes

(scales) $\mathcal{W}_k(x) = \mathrm{diag}(w_{k,1}, \ldots, w_{k,n})$ are nested. One can say that the localizing schemes are nested in the sense that for the corresponding matrices the following *ordering condition* is fulfilled:

(A3) *For any fixed x and the method of localization $w_{(\cdot,\cdot)}(x)$ the following relation holds:*

$$\mathcal{W}_1(x) \le \ldots \le \mathcal{W}_k(x) \le \ldots \le \mathcal{W}_K(x).$$

For the kernel smoothing this condition means the following. Let the sequence of bandwidths $\{h_k\}$ be ordered from the smallest to the largest one, i.e., $h_1 < \ldots < h_K$, and let $\mathcal{W}_k(x) = \mathrm{diag}(w_{k,1}, \ldots, w_{k,n})$ be the localizing matrix, corresponding to the bandwidth h_k. Here the weights $w_{k,i} = w_{k,i}(x) = W((X_i - x)/h_k) \in [0,1]$ are nonnegative functions such that $W(u/h_l) \le W(u/h_k)$ for any $0 < h_l < h_k < 1$, and $W(u) \to 0$ as $|u| \to \infty$, or even is compactly supported.

Recall that, given $x \in \mathcal{X}$, a basis $\{\psi_j\}$, and the method of localization $w_{(\cdot,\cdot)}(x)$, we look for the estimator $\widehat{f}_{\widehat{\theta}}(x)$ of $f(x)$ having form (1.5), where the coefficients $\widehat{\theta}^{(j)}(x)$ are the components of the estimator

$$\widehat{\boldsymbol{\theta}}(x) \stackrel{\mathrm{def}}{=} \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x) = (\widetilde{\theta}^{(1)}_{\widehat{k}}(x), \ldots, \widetilde{\theta}^{(p)}_{\widehat{k}}(x))^\top, \tag{2.10}$$

corresponding to the adaptive choice of the index $\widehat{k} \in \{1, \ldots, K\}$, i.e. to the choice of the degree of localization.

The selection of $\widehat{\boldsymbol{\theta}}(x)$ from $\{\widetilde{\boldsymbol{\theta}}_k(x)\}$, $k = 1, \ldots, K$ can be done by the application of the Lepski [20] method to the comparing of the maximized log-likelihoods $\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k)$. This is the idea of the FLL technique suggested in [18]. More precisely, to describe the test statistic, define for any $\boldsymbol{\theta}$, $\boldsymbol{\theta}' \in \Theta$ the corresponding log-likelihood ratio:

$$\mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}, \boldsymbol{\theta}') \stackrel{\mathrm{def}}{=} \mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}) - \mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}'). \tag{2.11}$$

Then, using the approach suggested in [18], for every $l = 1, \ldots, K$, *the fitted log-likelihood (FLL)* ratio is defined as follows:

$$\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \boldsymbol{\theta}') \stackrel{\mathrm{def}}{=} \max_{\boldsymbol{\theta} \in \Theta} \mathrm{L}(\mathbf{W}_l, \boldsymbol{\theta}, \boldsymbol{\theta}').$$

By Theorem 4.2, for any l and $\boldsymbol{\theta}$, the FLL is a quadratic form:

$$2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \boldsymbol{\theta}) = (\widetilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta})^\top \mathbf{B}_l (\widetilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}).$$

This prompts the use, see [18], the *FLL-statistics*:

$$\begin{aligned} T_{lk} &\stackrel{\mathrm{def}}{=} 2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_k) \\ &= (\widetilde{\boldsymbol{\theta}}_l - \widetilde{\boldsymbol{\theta}}_k)^\top \mathbf{B}_l (\widetilde{\boldsymbol{\theta}}_l - \widetilde{\boldsymbol{\theta}}_k)\,, \quad l < k. \end{aligned} \tag{2.12}$$

In the algorithm the smallest bandwidths corresponding to $k = 1$ is always accepted, then the adaptive index $\widehat{k}$ is selected by Lepski's selection rule with the FLL test statistics $\{T_{lm}\}$, $1 \le l < m \le K$:

$$\widehat{k} = \max\{k \le K : T_{lm} \le \mathfrak{z}_l,\, l < m \le k\}. \tag{2.13}$$

Finally we set $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{k}}$.

The procedure (2.13) involves parameters $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ related to the large deviations of $\{T_{lm}\}$, $1 \le l < m \le K$. As the classical Lepski procedure, the (2.13) controls the risk of estimators for the case of dominating bias. The opposite case of the negligible w.r.t. the noise bias is usually

handled by employing the advanced empirical process technique, however sometimes providing the constants far away of being optimal. Notice also that the Wilks-type Theorem 4.3 below gives the bound for the expected fitted log-likelihood ratio:

$$\mathbb{E}|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*)|^r \le C(p,r) \tag{2.14}$$

where the constant $C(p,r)$ does not depend on the degree of localization and is given by the r th moment of the χ^2 distribution with p degrees of freedom:

$$C(p,r) = \mathbb{E}|\chi_p^2|^r = 2^r\Gamma(r+p/2)/\Gamma(p/2), \tag{2.15}$$

Therefore we shall follow the practical idea from [27] and [18] allowing to avoid hard large deviations analysis and to calculate the thresholds rather sharp numerically. We assume at this step that the critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ are already fixed satisfying the following set of $K-1$ inequalities:

Definition 2.1. ***Propagation conditions (PC)***
Let $\widehat{\boldsymbol{\theta}}_k$ denote the last accepted estimate after the first k steps of the procedure:

$$\widehat{\boldsymbol{\theta}}_k \stackrel{\mathrm{def}}{=} \widetilde{\boldsymbol{\theta}}_{\min\{k,\widehat{k}\}}. \tag{2.16}$$

The critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ satisfy

$$\mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^r \le \alpha C(p,r) \quad \textit{for all } k = 2, \ldots, K, \tag{2.17}$$

where $C(p,r)$ is defined by (2.15), $\alpha \in (0,1]$ is an additional "free" turning parameter which can be taken equal to 1, and $\mathbb{E}_{0,\Sigma}$ stands for the expectation w.r.t. the measure $\mathcal{N}(0,\Sigma)$.

Remark 2.1. Lemma 4.1 from Section 4 shows that in the "no bias" situation the Gaussian distribution provides a nice pivotality property: the actual value of the parameter $\boldsymbol{\theta}$ is not important for the risk of adaptive estimate, so one can put $\boldsymbol{\theta} = 0$ in (2.17).

Remark 2.2. Clearly at any step $k \le K$ of the algorithm the "current value" of the adaptive estimator $\widehat{\boldsymbol{\theta}}_k$ depends on the thresholds $\mathfrak{z}_1, \ldots, \mathfrak{z}_{k-1}$. The theoretical aspects related to the heteroscedasticity of model and to the incorrectly known variance is the focus of the present paper. Thus we do not detail the practical aspects of the thresholds calibration only mentioning that in practise this can be done by Monte Carlo simulations under the known "parametric" measure $\mathcal{N}(0,\Sigma)$. Moreover one needs to calculate them only once. For detailed consideration of the practical aspects of the calibration as well as for the computational results see [27] or [18] focused on the image denoising by local constant fitting. Demo-versions of the software are available on the web page http://www.cs.tut.fi/~ lasip/.

3 Theoretical study

In order to control the admissible level of misspecification for "model" covariance matrix from (1.6) we need to introduce the following condition on the relative variability in errors:

(A4) *There exists $\delta \in [0,1)$ such that*

$$1 - \delta \le \sigma_{0,i}^2/\sigma_i^2 \le 1 + \delta \quad \textit{for all} \quad i = 1, \ldots, n.$$

3.1 Upper bound for the critical values

Let us at this step recall the notion of the *Löwner partial ordering*: for any real symmetric matrices A and B we write $A \preceq B$ if and only if $\vartheta^\top A\,\vartheta \le \vartheta^\top B\,\vartheta$ for all vectors ϑ, or, equivalently, if and only if the matrix $B - A$ is nonnegative definite. Assuming $(A4)$, the true covariance matrix $\Sigma_0 \preceq \Sigma(1+\delta)$, and the variance of the estimate $\widetilde{\boldsymbol{\theta}}_k$ is bounded with $\mathbf{B}_k^{-1}$:

$$\begin{aligned} V_k \stackrel{\text{def}}{=} \operatorname{Var}\widetilde{\boldsymbol{\theta}}_k &= \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0\mathbf{W}_k\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1} \qquad (3.1)\\ &\preceq (1+\delta)\mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma\mathbf{W}_k\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1}\\ &= (1+\delta)\mathbf{B}_k^{-1}\boldsymbol{\Psi}\Sigma^{-1/2}\mathcal{W}_k^2\Sigma^{-1/2}\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1}\\ &\preceq (1+\delta)\mathbf{B}_k^{-1}\boldsymbol{\Psi}\Sigma^{-1/2}\mathcal{W}_k\Sigma^{-1/2}\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1}\\ &= (1+\delta)\mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1}\\ &= (1+\delta)\mathbf{B}_k^{-1}. \qquad (3.2) \end{aligned}$$

The last inequality follows from the observation that all the entries of the "weight" matrix $\mathcal{W}_k$ do not exceed one, implying $\mathcal{W}_k^2 \preceq \mathcal{W}_k$. The strict equality takes place if the $\{w_{k,i}\}$ are boxcar (rectangular) kernels and the noise is known, i.e., if $\delta = 0$. To justify the procedure one need to show that the critical values chosen by (PC) are finite. This is obtained under the following assumption:

$(A5)$ *Let for some constants u_0 and u such that $1 < u_0 \le u$ for any $2 \le k \le K$ the matrices $\mathbf{B}_k$ satisfy*

$$u_0 I_p \preceq \mathbf{B}_{k-1}^{-1/2}\,\mathbf{B}_k\,\mathbf{B}_{k-1}^{-1/2} \preceq u I_p$$

Remark 3.1. In the "one dimensional case" $p = 1$, that is for the local constant fit, the "matrix" $\mathbf{B}_k = \sum_{i=1}^n w_{k,i}\sigma_i^{-2} \ge \mathbf{B}_{k-1}$ is just a weighted "local design size". Assume for simplicity that $\sigma_i^2 \equiv \sigma^2$, the weights are rectangular kernels $w_{k,i}(x) = \mathbb{I}\{|X_i - x| \le h_k/2\}$, and the design is equidistant. Then for n sufficiently large

$$\frac{1}{n}\mathbf{B}_k = \frac{1}{n\sigma^2}\sum_{i=1}^n \mathbb{I}\{|\frac{i}{n} - x| \le \frac{h_k}{2}\} \approx \frac{h_k}{\sigma^2},$$

and the condition $(A5)$ means that the bandwidths grow geometrically: $h_k = u h_{k-1}$.

Now we are able to formulate a theorem on finiteness of the critical values.

Theorem 3.1. ***The theoretical choice of the critical values***

Assume $(A1) - (A3)$ and $(A5)$. The adaptive procedure (2.13) in the considered set-up is well defined in the sense that the choice of the critical values of the form

$$\mathfrak{z}_k = \frac{4}{\mu}\left\{ r(K-k)\log u + \log(K/\alpha) - \frac{p}{4}\log(1-4\mu) - \log(1-u^{-r}) + \overline{C}(p,r)\right\} \qquad (3.3)$$

provides the conditions (2.17) for all $k \le K$. Here

$$\overline{C}(p,r) = \log\left\{ \frac{2^{2r}[\Gamma(2r+p/2)\Gamma(p/2)]^{1/2}}{\Gamma(r+p/2)}\right\}$$

and $\mu \in (0, 1/4)$. *Particularly,*

$$\mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_K - \widehat{\boldsymbol{\theta}})^\top \mathbf{B}_K (\widetilde{\boldsymbol{\theta}}_K - \widehat{\boldsymbol{\theta}})|^r \le \alpha C(p, r). \tag{3.4}$$

The proof is given in subsection 4.2.

3.2 Quality of estimation in the nearly parametric case: The small modeling bias condition and the propagation property

The critical values of the procedure $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ were selected by the propagation conditions (2.17) under the measure $\mathcal{N}(\boldsymbol{\theta}, \Sigma)$. Now $\boldsymbol{\theta}_1^* \approx \cdots \boldsymbol{\theta}_k^* \approx \boldsymbol{\theta}$ up to some $k \le K$, and the covariance matrix is Σ_0 . The aim is to formalize the meaning of "$\approx$" and to justify the use of the critical values in this situation. For this purposes we will take into account the discrepancy between the joint distributions of linear estimates $\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_k$ for $k = 1, \ldots, K$ under "no bias" assumption corresponding to the distributions with the mean $\boldsymbol{\theta}_1^* = \cdots = \boldsymbol{\theta}_k^* = \boldsymbol{\theta}$ and the incorrectly specified covariance matrix Σ , and in the general situation with $\boldsymbol{\theta}_1^* \ne \cdots \ne \boldsymbol{\theta}_k^*$ and the covariance Σ_0 . Denote the expectations w.r.t. these measures by $\mathbb{E}_{\boldsymbol{\theta},\Sigma} := \mathbb{E}_{k,\boldsymbol{\theta},\Sigma}$ and $\mathbb{E}_{\boldsymbol{f},\Sigma_0} := \mathbb{E}_{k,\boldsymbol{f},\Sigma_0}$, respectively. Denote a $p \times k$ matrix of the first k estimators and the expectations correspondingly by

$$\begin{aligned}
\widetilde{\boldsymbol{\Theta}}_k &\stackrel{\text{def}}{=} (\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_k), \\
\boldsymbol{\Theta}_k^* &\stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{f},\Sigma_0} \widetilde{\boldsymbol{\Theta}}_k = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*), \\
\boldsymbol{\Theta}_k &\stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{\theta},\Sigma} \widetilde{\boldsymbol{\Theta}}_k = (\boldsymbol{\theta}, \ldots, \boldsymbol{\theta}).
\end{aligned}$$

Let $A \otimes B$ stands for the Kronecker product of A and B defined as

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}.$$

Denote the $pk \times pk$ covariance matrices of $\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_k^\top = (\widetilde{\boldsymbol{\theta}}_1^\top, \ldots, \widetilde{\boldsymbol{\theta}}_k^\top) \in \mathbb{R}^{pk}$ by

$$\boldsymbol{\Sigma}_k \stackrel{\text{def}}{=} \operatorname{Var}_{\boldsymbol{\theta},\Sigma}[\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_k] = \mathbf{D}_k (J_k \otimes \Sigma) \mathbf{D}_k^\top, \tag{3.5}$$

$$\boldsymbol{\Sigma}_{k,0} \stackrel{\text{def}}{=} \operatorname{Var}_{\boldsymbol{f},\Sigma_0}[\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_k] = \mathbf{D}_k (J_k \otimes \Sigma_0) \mathbf{D}_k^\top, \tag{3.6}$$

where the matrix J_k is a $k \times k$ matrix with all its elements equal to 1 , and the $pk \times nk$ matrix $\mathbf{D}_k$ is defined as follows:

$$\begin{aligned}
\mathbf{D}_k &\stackrel{\text{def}}{=} D_1 \oplus \cdots \oplus D_k = \operatorname{diag}(D_1, \ldots, D_k), \\
D_l &\stackrel{\text{def}}{=} \mathbf{B}_l^{-1} \boldsymbol{\Psi} \mathbf{W}_l, \quad l = 1, \ldots, k.
\end{aligned} \tag{3.7}$$

By Lemma 4.8 from Section 4 under Assumption $(A4)$ with the same δ the similar relation holds for the covariance matrices $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Sigma}_{k,0}$ of the linear estimators:

$$(1 - \delta)\boldsymbol{\Sigma}_k \preceq \boldsymbol{\Sigma}_{k,0} \preceq (1 + \delta)\boldsymbol{\Sigma}_k \ , \quad k \le K. \tag{3.8}$$

In spite of the moment generating function of $\operatorname{vec}\widetilde{\boldsymbol{\Theta}}_K$ has the form corresponding to the multivariate normal distribution, see Lemma 4.10 in Section 4, this representation makes sense only if $\boldsymbol{\Sigma}_K$ is nonsingular. Notice that $\operatorname{rank}(J_K \otimes \Sigma) = n$. From $J_K \otimes \Sigma \succeq 0$ it follows only that $\boldsymbol{\Sigma}_K \succeq 0$, similarly, $\boldsymbol{\Sigma}_{K,0} \succeq 0$. However, without any additional assumptions it is easy to show, see Lemma 4.9 in Section 4, that for rectangular kernels $\boldsymbol{\Sigma}_K \succ 0$. On the other hand, due to (3.8), it is enough to require nonsingularity only for the matrix $\boldsymbol{\Sigma}_K$ corresponding to the approximate model (1.6), and its choice belongs to a statistician. In what follows we assume that $\boldsymbol{\Sigma}_K \succ 0$.

Denote by $I\!P^k_{\boldsymbol{\theta},\Sigma} = \mathcal{N}(\operatorname{vec}\boldsymbol{\Theta}_k, \boldsymbol{\Sigma}_k)$ and by $I\!P^k_{\boldsymbol{f},\Sigma_0} = \mathcal{N}(\operatorname{vec}\boldsymbol{\Theta}^*_k, \boldsymbol{\Sigma}_{k,0})$, $k = 1, \ldots, K$, the distributions of $\operatorname{vec}\widetilde{\boldsymbol{\Theta}}_k$ under the null and under the alternative. Denote also the Radon-Nikodym derivative by

$$Z_k \stackrel{\text{def}}{=} \frac{\mathrm{d}I\!P^k_{\boldsymbol{f},\Sigma_0}}{\mathrm{d}I\!P^k_{\boldsymbol{\theta},\Sigma}}. \tag{3.9}$$

Then, by Lemma 4.11 from Section 4, the Kullback-Leibler divergence between these measures has the following form:

$$\begin{aligned} 2\mathbb{KL}(I\!P^k_{\boldsymbol{f},\Sigma_0}, I\!P^k_{\boldsymbol{\theta},\Sigma}) &\stackrel{\text{def}}{=} 2\mathbb{E}_{\boldsymbol{f},\Sigma_0} \log\big(\frac{\mathrm{d}I\!P^k_{\boldsymbol{f},\Sigma_0}}{\mathrm{d}I\!P^k_{\boldsymbol{\theta},\Sigma}}\big) \\ = \quad \Delta(k) + \log\left(\frac{\det\boldsymbol{\Sigma}_k}{\det\boldsymbol{\Sigma}_{k,0}}\right) &+ \operatorname{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{k,0}) - pk, \end{aligned} \tag{3.10}$$

where

$$b(k) \stackrel{\text{def}}{=} \operatorname{vec}\boldsymbol{\Theta}^*_k - \operatorname{vec}\boldsymbol{\Theta}_k \tag{3.11}$$

$$\Delta(k) \stackrel{\text{def}}{=} b(k)^\top \boldsymbol{\Sigma}_k^{-1} b(k). \tag{3.12}$$

If there would be no any "noise misspecification", i.e., if $\delta \equiv 0$ implying $\Sigma = \Sigma_0$, then $\Delta(k) = b(k)^\top \boldsymbol{\Sigma}_k^{-1} b(k) = 2\mathbb{KL}(I\!P^k_{\boldsymbol{f},\Sigma}, I\!P^k_{\boldsymbol{\theta},\Sigma})$. Therefore, this quantity can be used to indicate deviation between the mean values in the true (1.1) and the approximate (1.6) models. Clearly, under $(\mathcal{W})$, the quantity $\Delta(k)$ grows with k, so following the terminology suggested in [27], we introduce the *small modeling bias condition*:

(SMB) *Let for some $k \le K$ and some $\boldsymbol{\theta}$ exist a constant $\Delta \ge 0$ such that*

$$\Delta(k) \le \Delta.$$

Monotonicity of $\Delta(k)$ and Assumption (SMB) immediately imply that

$$\Delta(k') \le \Delta \text{ for all } k' \le k.$$

The conditions (3.8) yield $-pk\delta \le \operatorname{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{k,0}) - pk \le pk\delta$. Thus (4.19) implies the bound for the Kullback-Leibler divergence in terms of δ:

$$-\frac{pk}{2}\log(1+\delta) + \frac{\Delta(k)}{2} - \frac{pk\delta}{2} \le \mathbb{KL}(I\!P^k_{\boldsymbol{f},\Sigma_0}, I\!P^k_{\boldsymbol{\theta},\Sigma}) \le -\frac{pk}{2}\log(1-\delta) + \frac{\Delta(k)}{2} + \frac{pk\delta}{2}. \tag{3.13}$$

Moreover, as $\delta \to 0+$

$$\Delta(k) - 2pk\delta + o(\delta) \le 2\mathbb{KL}(I\!P^k_{\boldsymbol{f},\Sigma_0}, I\!P^k_{\boldsymbol{\theta},\Sigma}) \le \Delta(k) + 2pk\delta + o(\delta). \tag{3.14}$$

This means that, if for some k Assumption (SMB) is fulfilled and $\delta = O(1/K)$, then the Kullback-Leibler divergence between the measures $I\!P^k_{\boldsymbol{\theta},\Sigma}$ and $I\!P^k_{\boldsymbol{f},\Sigma_0}$ is bounded by a small constant.

Now one can state the crucial property for obtaining the final oracle result.

Theorem 3.2. ***Propagation property***

Assume $(A1)-(A5)$ *and* (PC). *Then for any* $k \le K$ *the following upper bounds hold:*

$$\begin{aligned} & \mathbb{E}|(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})|^{r/2} \\ \le\ & (\mathbb{E}|\chi^2_p|^r)^{1/2} (1+\delta)^{pk/4} (1-\delta)^{-3pk/4} \exp\left\{ \varphi(\delta) \frac{\Delta(k)}{2(1-\delta)} \right\}, \\ & \mathbb{E}|(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^{r/2} \\ \le\ & (\alpha \mathbb{E}|\chi^2_p|^r)^{1/2} (1+\delta)^{pk/4} (1-\delta)^{-3pk/4} \exp\left\{ \varphi(\delta) \frac{\Delta(k)}{2(1-\delta)} \right\}, \end{aligned}$$

where $\varphi(\delta) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{for homogeneous errors,} \\ \frac{2(1+\delta)}{(1-\delta)^2} - 1 & \text{otherwise.} \end{cases}$

Here $\widetilde{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_k(x)$ *is the QMLE defined by* (2.2), *and* $\widehat{\boldsymbol{\theta}}_k(x) = \widetilde{\boldsymbol{\theta}}_{\min k, \widehat{k}}(x)$ *is the adaptive estimate at the* k *th step of the procedure.*

The proof is given in Subsection 4.4.

Remark 3.2. Bounds (4.28) and (4.27) below give a condition on the relative error in the noise misspecification. As $\delta \to 0+$ for every $k \le K$ it holds that

$$\varphi(\delta) \frac{\Delta(k)}{1+\delta} - 2pk\delta + o(\delta) \le \log \mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2] \le \varphi(\delta) \frac{\Delta(k)}{1-\delta} + 2pk\delta + o(\delta),$$

where Z_k is defined by (3.9). This bound implies, up to the additive constant $\log\bigl(\alpha \mathbb{E}|\chi^2_p|^r\bigr)/2$, the same asymptotic behavior for the logarithm of the risk of adaptive estimate at each step of the procedure. Because by (SMB) the quantity $\Delta(k)$ is supposed to be bounden by a small constant, and K is of order $\log n$, then $\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2]$ is small if $\delta = O(1/\log n)$. This means that for the case when Σ is an estimate for Σ_0 only the logarithmic in sample size quality is needed. This observation is of particular importance, since it is known from [26] that over classes of functions with bounded second derivative the rate $n^{-1/2}$ of variance estimation is achievable only for the dimension $d \le 8$.

Remark 3.3. The propagation property provides the adaptive procedure do not stop with high probability while $\Delta(k)$ is small, i.e., under (SMB), and if the relative error δ in the noise is sufficiently small.

3.3 Quality of estimation in the nonparametric case: the oracle result

Define the *oracle index* as the largest index $k \le K$ such that (SMB) holds:

$$k^* \stackrel{\text{def}}{=} \max\{k \le K : \Delta(k) \le \Delta\}. \tag{3.15}$$

Theorem 3.3. *Let $\Delta(1) \leq \Delta$, i.e., the first estimate is always accepted in the testing procedure. Let k^* be the oracle index. Then under the conditions $(A1)$, $(A2)$, $(A4)$, $(\mathcal{W})$, $(A5)$ the risk between the adaptive estimate and the oracle one is bounded with the following expression:*

$$\begin{aligned} & \mathbb{E}|(\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})|^{r/2} \\ \leq \;& \mathfrak{z}_{k^*}^{r/2} + (\alpha \mathbb{E}|\chi_p^2|^r)^{1/2} (1+\delta)^{pk^*/4} (1-\delta)^{-3pk^*/4} \exp\left\{ \varphi(\delta) \frac{\Delta}{2(1-\delta)} \right\}, \end{aligned} \tag{3.16}$$

where $\varphi(\delta)$ is as in Theorem 3.2.

Proof. By the definition of the adaptive estimate $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{k}}$. Because the events $\{\widehat{k} \leq k^*\}$ and $\{\widehat{k} > k^*\}$ are disjunct, one can write

$$\begin{aligned} & \mathbb{E}|(\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})|^{r/2} \\ = \;& \mathbb{E}|(\widetilde{\boldsymbol{\theta}}_{k^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{k}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{k}})|^{r/2} \mathbb{I}\{\widehat{k} \leq k^*\} \\ + \;& \mathbb{E}|(\widetilde{\boldsymbol{\theta}}_{k^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{k}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{k}})|^{r/2} \mathbb{I}\{\widehat{k} > k^*\}. \end{aligned}$$

If $\widehat{k} \leq k^*$ then $\widehat{\boldsymbol{\theta}}_{k^*} \stackrel{\text{def}}{=} \widetilde{\boldsymbol{\theta}}_{\min\{k^*, \widehat{k}\}} = \widetilde{\boldsymbol{\theta}}_{\widehat{k}}$. Thus, to bound the first summand, it is enough to apply Theorem 3.2 with $k = k^*$.

To bound the second expectation, i.e., to bound fluctuations of the adaptive estimate $\widehat{\boldsymbol{\theta}}$ at the steps of the procedure for which the (SMB) condition is not fulfilled anymore, just notice that for $\widehat{k} > k^*$ the quadratic form coincides with the test statistics $T_{k^*, \widehat{k}}$

$$\begin{aligned} & (\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}}) \\ = \;& (\widetilde{\boldsymbol{\theta}}_{k^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{k}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widetilde{\boldsymbol{\theta}}_{\widehat{k}}) \stackrel{\text{def}}{=} T_{k^*, \widehat{k}}. \end{aligned}$$

But the index $\widehat{k}$ was accepted, this means that $T_{l, \widehat{k}} \leq \mathfrak{z}_l$ for all $l < \widehat{k}$ and therefore for $l = k^*$. Thus

$$\mathbb{E}|(\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})^\top \mathbf{B}_{k^*} (\widetilde{\boldsymbol{\theta}}_{k^*} - \widehat{\boldsymbol{\theta}})|^{r/2} \mathbb{I}\{\widehat{k} > k^*\} \leq \mathfrak{z}_{k^*}^{r/2}.$$

□

3.4 Componentwise oracle risk bounds

Theorem 3.3 provides the oracle risk bound for the adaptive estimator $\widehat{\boldsymbol{\theta}}(x) = \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)$ of the parameter vector $\boldsymbol{\theta}(x) \in \mathbb{R}^p$ corresponding to the estimator $\widehat{f}_{\widehat{\boldsymbol{\theta}}}(x)$ of the type (1.5). It is interesting to have a look at the oracle quality of estimation of the components $\theta^{(1)}, \ldots, \theta^{(p)}$ of the vector $\boldsymbol{\theta}$ having in mind that the choice of polynomial basis leads to the direct estimation of the value of regression function and the derivatives by the coordinates of $\widehat{\boldsymbol{\theta}}$.

Denote by $LP_k(p-1)$ a local polynomial estimator of order $p-1$ corresponding to the kth degree of localization, and, respectively, by $LP^{ad}(p-1)$ its adaptive counterpart, i.e., $LP^{ad}(p-1) \stackrel{\text{def}}{=} LP_{\widehat{k}}(p-1)$. If the basis is polynomial and the regression function $f(\cdot)$ is sufficiently smooth in a neighborhood of x, then $\widehat{\boldsymbol{\theta}}(x)$ is the $LP^{ad}(p-1)$ of the vector $(f^{(0)}(x), \ldots, f^{(p-1)}(x))^\top$ of the values of the function f and its derivatives at the reference point $x \in \mathbb{R}^d$.

Now we are going to obtain a similar to the previous section oracle result for the components of the vector $\widehat{\boldsymbol{\theta}}(x)$, particularly for $\boldsymbol{e}_j^\top\widehat{\boldsymbol{\theta}}(x)$, $j=1,\ldots,p$, where $\boldsymbol{e}_j=(0,\ldots,1,\ldots,0)^\top$ is the jth canonical basis vector in $\mathbb{R}^p$. As a corollary of this general result in the case of the polynomial basis we get an oracle risk bound for $LP^{ad}(p-1)$ estimators of the function f and its derivatives at the point x.

Let $LP_k(p-1)$ estimator of $f^{(j-1)}(x)$ be given by

$$\begin{aligned}\widetilde{f}_k^{(j-1)}(x) &= e_j^\top\widetilde{\boldsymbol{\theta}}_k(x),\ j=1,\ldots,p, \\ \widetilde{f}_k(x) &= \widetilde{f}_k^{(0)}(x)=e_1^\top\widetilde{\boldsymbol{\theta}}_k(x).\end{aligned} \tag{3.17}$$

Then the adaptive local polynomial estimators are defined as follows:

$$\begin{aligned}\widehat{f}^{(j-1)}(x) &= e_j^\top\widehat{\boldsymbol{\theta}}(x),\ j=1,\ldots,p, \\ \widehat{f}(x) &= e_1^\top\widehat{\boldsymbol{\theta}}(x).\end{aligned} \tag{3.18}$$

Similarly, the adaptive estimators of the function f and its derivatives corresponding to the kth step of the procedure are given by

$$\widehat{f}_k^{(j-1)}(x)\stackrel{\mathrm{def}}{=}e_j^\top\widehat{\boldsymbol{\theta}}_k(x),\ j=1,\ldots,p. \tag{3.19}$$

Thus, if the basis is polynomial, the estimator $\widehat{f}(x)\stackrel{\mathrm{def}}{=}\widehat{f}^{(0)}(x)$ is the $LP^{ad}(p-1)$ estimator of the value $f(x)$, and $\widehat{f}^{(j-1)}(x)$ with $j=2,\ldots,p$ are, correspondingly, the $LP^{ad}(p-1)$ estimators of the values of its derivatives. However it should be stressed that the results of Theorems 3.3 and 3.9 hold for any basis satisfying the conditions of the theorems. We shall need the following assumptions:

(A6) *There exist $0<\sigma_{min}(k)\le\sigma_{max}(k)<\infty$ such that for $i: X_i\in U_{h_k}(x)$, with $U_{h_k}(x)$ given by $\mathbf{W}_k$ the variances of errors from the parametric (known) model* (1.6) *are locally uniformly bounded:*

$$\sigma^2_{min}(k)\le\sigma_i^2\le\sigma^2_{max}(k).$$

(A7) *Let assumption* (A6) *be satisfied. There exists a number $\Lambda_0>0$ such that for any $k=1,\ldots,K$ the smallest eigenvalue $\lambda_p(\mathbf{B}_k)\ge nh_k^d\Lambda_0\sigma^{-2}_{max}(k)$ for n sufficiently large.*

Then, because $\mathbf{B}_k\succ 0$, for any $k=1,\ldots K$ and for any $\gamma\in\mathbb{R}^p$ we have

$$\gamma^\top\mathbf{B}_k^{-1}\gamma\le\frac{\sigma^2_{max}(k)}{nh_k^d\Lambda_0}\|\gamma\|^2\le\frac{\overline{\sigma}^2_{max}(k)}{nh_k^d\Lambda_0}\|\gamma\|^2, \tag{3.20}$$

where $\overline{\sigma}^2_{max}(k)\stackrel{\mathrm{def}}{=}\max_{1\le l\le k}\sigma^2_{max}(l)$. Thus we have the following lemma:

Lemma 3.4. *Let* (A6) *and* (A7) *be satisfied. Then for any $j=1,\ldots,p$ and $k,k'=1,\ldots K$ the following upper bound holds:*

$$\left(\frac{nh_k^d\Lambda_0}{\overline{\sigma}^2_{max}(k)}\right)^{1/2}|\boldsymbol{e}_j^\top\widetilde{\boldsymbol{\theta}}_k-\boldsymbol{e}_j^\top\widetilde{\boldsymbol{\theta}}_{k'}|\le\|\mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k-\widetilde{\boldsymbol{\theta}}_{k'})\|.$$

Proof. By (3.20) taking $\gamma = \mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_{k'})$ we have

$$\begin{aligned} |\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_k - \boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_{k'}|^2 &\leq \|\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_{k'}\|^2 \\ &= \|\mathbf{B}_k^{-1/2}\mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_{k'})\|^2 \\ &\leq \frac{\overline{\sigma}^2_{max}(k)}{nh_k^d\Lambda_0}\|\mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_{k'})\|^2. \end{aligned}$$

□

To obtain the "componentwise" oracle risk bounds we need to recheck the "propagation property". Firstly, notice that the "propagation conditions" (2.17) on the choice the critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ imply the similar bounds for the components $\boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}_k(x)$. Recall that $\widehat{\boldsymbol{\theta}}_k \stackrel{\text{def}}{=} \widetilde{\boldsymbol{\theta}}_{\min\{k,\widehat{k}\}}$. Then, by (2.17), Lemma 3.4, and the pivotality property from Lemma 4.1, we have the following simple observation:

Lemma 3.5. *Under the propagation conditions (PC) for any $\boldsymbol{\theta} \in \mathbb{R}^p$ and all $k = 2, \ldots, K$ we have:*

$$\begin{aligned} \left(\frac{nh_k^d\Lambda_0}{\overline{\sigma}^2_{max}(k)}\right)^r \mathbb{E}_{\boldsymbol{\theta},\Sigma}|\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_k(x) - \boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}_k(x)|^{2r} &\leq \mathbb{E}_{0,\Sigma}\|\mathbf{B}_k^{1/2}(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)\|^{2r} \\ &\leq \alpha C(p,r). \end{aligned}$$

Here $\mathbb{E}_{0,\Sigma}$ stands for the expectation w.r.t. $\mathcal{N}(0,\Sigma)$ and $C(p,r)$ is given by (4.8).

As before we suppress the dependence on x. To get the propagation property we study for $k = 1, \ldots, K$ the joint distributions of $\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_1, \ldots, \boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_k$, that is the distribution of $\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\Theta}}_k$, the j th row of the matrix $\widetilde{\boldsymbol{\Theta}}_k$. Obviously,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{f},\Sigma_0}[\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\Theta}}_k] &= \boldsymbol{e}_j^\top \boldsymbol{\Theta}_k^* = (\boldsymbol{e}_j^\top \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{e}_j^\top \boldsymbol{\theta}_k^*), \\ \mathbb{E}_{\boldsymbol{\theta},\Sigma}[\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\Theta}}_k] &= \boldsymbol{e}_j^\top \boldsymbol{\Theta}_k = (\boldsymbol{e}_j^\top \boldsymbol{\theta}, \ldots, \boldsymbol{e}_j^\top \boldsymbol{\theta}). \end{aligned}$$

Recall that the matrices $\boldsymbol{\Sigma}_{k,0}$ and $\boldsymbol{\Sigma}_k$ have a block structure. Now, for instance, to study the estimator of the first coordinate of the vector $\boldsymbol{\theta} = \boldsymbol{\theta}(x)$, or of $f(x)$ in the case of the polynomial basis, we take the first elements of each block and so on. Denote the $k \times k$ covariance matrices of the j th elements of the vectors $\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_k$ by

$$\begin{aligned} \boldsymbol{\Sigma}_{k,j} &\stackrel{\text{def}}{=} \left\{ \operatorname{cov}_{\boldsymbol{\theta},\Sigma}\left[\widetilde{\theta}_l^{(j)}, \widetilde{\theta}_m^{(j)}\right]\right\}_{1\leq l\leq m\leq k} \\ &= \mathbf{D}_{k,j}(J_k \otimes \Sigma)\mathbf{D}_{k,j}^\top, \qquad (3.21) \\ \boldsymbol{\Sigma}_{k,0,j} &\stackrel{\text{def}}{=} \left\{ \operatorname{cov}_{\boldsymbol{f},\Sigma_0}\left[\widetilde{\theta}_l^{(j)}, \widetilde{\theta}_m^{(j)}\right]\right\}_{1\leq l\leq m\leq k} \\ &= \mathbf{D}_{k,j}(J_k \otimes \Sigma_0)\mathbf{D}_{k,j}^\top, \qquad (3.22) \end{aligned}$$

where J_k is a $k \times k$ matrix with all its elements equal to 1, and the $k \times nk$ block diagonal matrices $\mathbf{D}_{k,j}$ is defined by

$$\begin{aligned} \mathbf{D}_{k,j} &\stackrel{\text{def}}{=} \boldsymbol{e}_j^\top D_1 \oplus \cdots \oplus \boldsymbol{e}_j^\top D_k, = \left(I_k \otimes \boldsymbol{e}_j^\top\right)\mathbf{D}_k \\ D_l &\stackrel{\text{def}}{=} \mathbf{B}_l^{-1}\boldsymbol{\Psi}\mathbf{W}_l, \quad l = 1, \ldots, k. \qquad (3.23) \end{aligned}$$

Moreover, the following representation holds:

$$\begin{aligned}\boldsymbol{\Sigma}_{k,j} &= \bigl(I_k \otimes \boldsymbol{e}_j^\top\bigr)\mathbf{D}_k\bigl(J_k \otimes \Sigma\bigr)\mathbf{D}_k^\top\bigl(I_k \otimes \boldsymbol{e}_j^\top\bigr)^\top \\ &= \bigl(I_k \otimes \boldsymbol{e}_j\bigr)^\top \boldsymbol{\Sigma}_k \bigl(I_k \otimes \boldsymbol{e}_j\bigr), \end{aligned} \tag{3.24}$$

where $\boldsymbol{\Sigma}_k$ is defined by (3.5). Similarly,

$$\boldsymbol{\Sigma}_{k,0,j} = \bigl(I_k \otimes \boldsymbol{e}_j\bigr)^\top \boldsymbol{\Sigma}_{k,0} \bigl(I_k \otimes \boldsymbol{e}_j\bigr). \tag{3.25}$$

Thus, the important relation (3.8) is preserved for $\boldsymbol{\Sigma}_{k,j}$ and $\boldsymbol{\Sigma}_{k,0,j}$ obtained by picking up the (j,j) th elements of each block of $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Sigma}_{k,0}$ respectively.

With usual notation $\gamma^{(j)}$ for the j th component of $\gamma \in \mathbb{R}^k$, denote by

$$\begin{aligned} b_j(k) &\stackrel{\mathrm{def}}{=} (\boldsymbol{e}_j^\top(\boldsymbol{\theta}_1^* - \boldsymbol{\theta}), \ldots, \boldsymbol{e}_j^\top(\boldsymbol{\theta}_k^* - \boldsymbol{\theta}))^\top \\ &= ((\boldsymbol{\theta}_1^* - \boldsymbol{\theta})^{(j)}, \ldots, (\boldsymbol{\theta}_k^* - \boldsymbol{\theta})^{(j)})^\top \in \mathbb{R}^k \qquad (3.26) \\ \Delta_j(k) &\stackrel{\mathrm{def}}{=} b_j(k)^\top \boldsymbol{\Sigma}_{k,j}^{-1}\, b_j(k). \qquad (3.27) \end{aligned}$$

Theorem 3.6. ***"Componentwise" propagation property***

Under the conditions $(A1)-(A7)$ *and* (PC) *for any* $k \le K$ *the following upper bound holds:*

$$\begin{aligned} &\left(\frac{nh_k^d\Lambda_0}{\overline{\sigma}^2_{max}(k)}\right)^{r/2} \mathbb{E}|\boldsymbol{e}_j^\top\widetilde{\boldsymbol{\theta}}_k(x) - \boldsymbol{e}_j^\top\widehat{\boldsymbol{\theta}}_k(x)|^r \\ \le\ & (\alpha\mathbb{E}|\chi_p^2|^r)^{1/2}(1+\delta)^{pk/4}(1-\delta)^{-3pk/4}\exp\left\{\varphi(\delta)\frac{\Delta_j(k)}{2(1-\delta)}\right\} \end{aligned} \tag{3.28}$$

with $\varphi(\delta)$ *as in Theorem 3.2.*

Corollary 3.7. *Let the basis be polynomial. Then under the conditions of the preceding theorem* $\mathbb{E}|\widetilde{f}_k^{(j-1)}(x) - \widehat{f}_k^{(j-1)}(x)|^r$ *satisfy* (3.28)

Proof. The proof essentially follows the line of the proof of Theorem 3.2. If the distributions of $\operatorname{vec}\widetilde{\boldsymbol{\Theta}}_k$ were Gaussian, then any subvector is also Gaussian.

Denote by $I\!P_{\boldsymbol{\theta},\Sigma}^{k,j} = \mathcal{N}\bigl((\boldsymbol{e}_j^\top\boldsymbol{\theta}, \ldots, \boldsymbol{e}_j^\top\boldsymbol{\theta})^\top, \boldsymbol{\Sigma}_{k,j}\bigr)$ and by $I\!P_{\boldsymbol{f},\Sigma_0}^{k,j} = \mathcal{N}\bigl((\boldsymbol{e}_j^\top\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{e}_j^\top\boldsymbol{\theta}_k^*)^\top, \boldsymbol{\Sigma}_{k,0,j}\bigr)$, $k = 1, \ldots, K$, the distributions of $e_j^\top\widetilde{\boldsymbol{\Theta}}_k$ under the null and under the alternative.

By the Cauchy-Schwarz inequality and Lemma 3.5

$$\left(\frac{nh_k^d\Lambda_0}{\overline{\sigma}^2_{max}(k)}\right)^{r/2} \mathbb{E}|\boldsymbol{e}_j^\top\widetilde{\boldsymbol{\theta}}_k(x) - \boldsymbol{e}_j^\top\widehat{\boldsymbol{\theta}}(x)|^r \le (\alpha\mathbb{E}|\chi_p^2|^r)^{1/2}\bigl(\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_{k,j}^2]\bigr)^{1/2}$$

with the Radon-Nikodym derivative given by $Z_{k,j} = \mathrm{d}I\!P_{\boldsymbol{f},\Sigma_0}^{k,j}/\mathrm{d}I\!P_{\boldsymbol{\theta},\Sigma}^{k,j}$. By inequalities (3.24) and (3.25) the analog of Condition $(A4)$ is preserved for $\boldsymbol{\Sigma}_{k,0,j}$ and $\boldsymbol{\Sigma}_{k,j}$, that is, there exist $\delta \in [0,1)$ such that

$$(1-\delta)\boldsymbol{\Sigma}_{k,j} \preceq \boldsymbol{\Sigma}_{k,0,j} \preceq (1+\delta)\boldsymbol{\Sigma}_{k,j} \tag{3.29}$$

for any $k \le K$ and $j = 1, \ldots, p$. Then the assertion of the theorem follows by the Taylor expansion at the point $(\boldsymbol{e}_j^\top\boldsymbol{\theta}, \ldots, \boldsymbol{e}_j^\top\boldsymbol{\theta})^\top$ and (3.29) similarly to the proof of Theorem 3.2. □

At this point we introduce the "componentwise" small modeling bias conditions:

(*SMBj*) *Let for some* $j = 1, \dots, p$, *some* $k(j) \le K$, *and some* $\theta^{(j)} = e_j^\top \boldsymbol{\theta}$ *exist a constant* $\Delta_j \ge 0$ *such that*

$$\Delta_j(k(j)) \le \Delta_j, \tag{3.30}$$

where $\Delta_j(k)$ *is defined by* (3.27).

Definition 3.8. *For each* $j = 1, \dots, p$ *the oracle index* $k^*(j)$ *is defined as the largest index in the scale for which the* $(SMBj)$ *condition holds, that is*

$$k^*(j) = \max\{k \le K : \Delta_j(k) \le \Delta_j\}. \tag{3.31}$$

Theorem 3.9. *Assume* $(A1) - (A7)$ *and* (PC). *Let the smallest bandwidth* h_1 *be such that the first estimate* $\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_1(x)$ *be always accepted in the adaptive procedure. Let* $k^*(j)$ *be the oracle index defined by* (3.31), $j = 1, \dots, p$. *Then the risk between the* j *th coordinates of the adaptive estimate and the oracle one is bounded with the following expression:*

$$\begin{aligned} &\left(\frac{nh_{k^*(j)}^d \Lambda_0}{\overline{\sigma}^2_{max}(k^*)}\right)^{r/2} \mathbb{E}|\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_{k^*(j)}(x) - \boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}(x)|^r \\ \le\ & \mathfrak{z}_{k^*(j)}^{r/2} + (\alpha \mathbb{E}|\chi_p^2|^r)^{1/2} (1+\delta)^{pk_j^*/4} (1-\delta)^{-3pk_j^*/4} \exp\left\{\varphi(\delta) \frac{\Delta_j}{2(1-\delta)}\right\}, \end{aligned} \tag{3.32}$$

where $\varphi(\delta)$ *is as in Theorem 3.6.*

Corollary 3.10. *Let the basis be polynomial. Then under the conditions of the preceding theorem the risk between the adaptive estimate and the oracle one* $\mathbb{E}|\widetilde{f}^{(j-1)}_{k^*(j)}(x) - \widehat{f}^{(j-1)}(x)|^r$ *satisfy* (3.32).

Proof. To simplify the notation we suppress the dependence on j in the index k. Similarly to the proof of Theorem 3.3 we consider the disjunct events $\{\widehat{k} \le k^*\}$ and $\{\widehat{k} > k^*\}$. Therefore,

$$\begin{aligned} & \mathbb{E}|\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_{k^*}(x) - \boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}(x)|^r \\ =\ & \mathbb{E}|\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_{k^*}(x) - \boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}(x)|^r \, \mathbb{I}\{\widehat{k} \le k^*\} \\ +\ & \mathbb{E}|\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_{k^*}(x) - \boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}(x)|^r \, \mathbb{I}\{\widehat{k} > k^*\}. \end{aligned}$$

By Lemma 3.4 and the definition of the test statistic $T_{k^*,\widehat{k}}$ the second summand can be easily bounded:

$$\begin{aligned} & \left(\frac{nh_{k^*}^d \Lambda_0}{\overline{\sigma}^2_{max}(k^*)}\right)^{r/2} \mathbb{E}|\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_{k^*}(x) - \boldsymbol{e}_j^\top \widehat{\boldsymbol{\theta}}(x)|^r \, \mathbb{I}\{\widehat{k} > k^*\} \\ \le\ & \mathbb{E}\|\mathbf{B}_{k^*}^{1/2}(\widetilde{\boldsymbol{\theta}}_{k^*}(x) - \widehat{\boldsymbol{\theta}}(x))\|^r \, \mathbb{I}\{\widehat{k} > k^*\} \\ \le\ & \mathfrak{z}_{k^*}^{r/2}. \end{aligned}$$

To bound the first summand we use the "componentwise" analog of Theorem 3.2, particularly, Theorem 3.6, and this completes the proof. □

3.5 SMB and the bias-variance trade-off

In [27] it was shown that the small modeling bias $(SMB1)$ condition (3.30) can be obtained from the "bias-variance trade-off" relations. Notice that our set-up includes the set-up from [27] as a particular case. To prove that the similar relation holds in the present case we need the following definition. Given a point x and the method of localization w, for any $j = 1, \ldots, p$ the "ideal adaptive bandwidths", see [22], [23] is defined as follows:

$$k^\star(j) = \max\{k \le K : \overline{b}_{k,f^{(j-1)}}(x) \le C_j(w)\sigma_k(x)\sqrt{d(n)}\}, \tag{3.33}$$

where $C_j(w)$ is a constant depending on the choice of the smoother w,

$$\begin{aligned} \overline{b}_{k,f^{(j-1)}}(x) &= \sup_{1\le l\le k} |\boldsymbol{e}_j^\top \boldsymbol{\theta}_l^*(x) - f^{(j-1)}(x)|, \\ \sigma_k^2(x) &= \mathrm{Var}_{\boldsymbol{f},\Sigma_0}[\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_k(x)], \\ d(n) &= \log(h_K/h_1), \end{aligned}$$

and $f^{(0)}$ stands for the function f itself. To bound the "modeling bias" $\Delta_j(k)$ we need the following assumption:

$(A8)$ *There exists a constant $s_j > 0$ such that for all $k \le K$*

$$\boldsymbol{\Sigma}_{k,j}^{-1} \preceq s_j \boldsymbol{\Sigma}_{k,j,diag}^{-1} \tag{3.34}$$

where $\boldsymbol{\Sigma}_{k,j,diag} = \mathrm{diag}\big(\mathrm{Var}_{\boldsymbol{\theta},\Sigma}[\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_1(x)], \ldots, \mathrm{Var}_{\boldsymbol{\theta},\Sigma}[\boldsymbol{e}_j^\top \widetilde{\boldsymbol{\theta}}_k(x)]\big)$ is a diagonal matrix composed of the diagonal elements of $\boldsymbol{\Sigma}_{k,j}$. Thus we have the following result:

Theorem 3.11. *Assume $(A4)$, $(A5)$ and $(A8)$. Let the weights $\{w_{k,i}(x)\}$ satisfy (4.15). Then for any given point x, smoothing function w, and $j = 1, \ldots, p$, the choice of $k(j) = k^\star(j)$ defined by the relation (3.33) with $d(n) = 1$ implies the $(SMBj)$ condition $\Delta_j(k(j)) \le \Delta_j$ with the constant $\Delta_j = s_j C_j^2(w)(1+\delta)(1-u_0^{-1})^{-1}$.*

Proof. Consider the quantity $b_j(k)^\top \boldsymbol{\Sigma}_{k,j,diag}^{-1} b_j(k)$. Suppose that $\boldsymbol{e}_j^\top \boldsymbol{\theta}(x) = f^{(j-1)}(x)$. In view of relation (4.15) for the weights $\{w_{l,i}(x)\}$ the form of the matrix $\boldsymbol{\Sigma}_{k,j,diag}$ is particularly simple:

$$\boldsymbol{\Sigma}_{k,j,diag} = \mathrm{diag}(\boldsymbol{e}_j^\top \mathbf{B}_1^{-1} \boldsymbol{e}_j, \ldots, \boldsymbol{e}_j^\top \mathbf{B}_k^{-1} \boldsymbol{e}_j).$$

Then by $(A5)$ and (3.2)

$$\begin{aligned} b_j(k)^\top \boldsymbol{\Sigma}_{k,j,diag}^{-1} b_j(k) &= \sum_{l=1}^k \frac{|\boldsymbol{e}_j^\top(\boldsymbol{\theta}_l^* - \boldsymbol{\theta})|^2}{\boldsymbol{e}_j^\top \mathbf{B}_l^{-1} \boldsymbol{e}_j} \\ &\le \big(\overline{b}_{k,f^{(j-1)}}(x)\big)^2 \sum_{l=1}^k \frac{1}{\boldsymbol{e}_j^\top \mathbf{B}_l^{-1} \boldsymbol{e}_j} \\ &\le \frac{\big(\overline{b}_{k,f^{(j-1)}}(x)\big)^2}{\boldsymbol{e}_j^\top \mathbf{B}_k^{-1} \boldsymbol{e}_j} \sum_{l=1}^k u_0^{-(k-l)} \\ &\le \frac{\big(\overline{b}_{k,f^{(j-1)}}(x)\big)^2 (1+\delta)}{\sigma_k^2(x)(1-u_0^{-1})}. \end{aligned}$$

By (3.33) with $d(n) = 1$, the choice of $k = k^\star(j)$ implies $\left(\overline{b}_{k,f^{(j-1)}}(x)\right)^2 \le C_j^2(w)\sigma_k^2(x)$. Thus

$$b_j(k)^\top \mathbf{\Sigma}_{k,j,diag}^{-1} b_j(k) \le (1+\delta) C_j^2(w)(1-u_0^{-1})^{-1}$$

and

$$\Delta_j(k) = b_j(k)^\top \mathbf{\Sigma}_{k,j}^{-1} b_j(k) \le s_j C_j^2(w)(1+\delta)(1-u_0^{-1})^{-1}.$$

□

Remark 3.4. Using the standard technique it is easy to derive from the above result that for estimation of functions over Hölder classes the methodology proposed in [18] and [27] and generalized in the present paper delivers the minimax rate of convergence up to a logarithmic factor.

4 Appendix

4.1 Pivotality and local parametric risk bounds

Lemma 4.1. ***Pivotality property***

Let (A3) hold. Let $\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_\varkappa = \boldsymbol{\theta}$ for $\varkappa \le K$. Then for any $k \le \varkappa$ the risk associated with the adaptive estimate at every step of the procedure does not depend on the parameter $\boldsymbol{\theta}$:

$$\mathbb{E}_{\boldsymbol{\theta}} |(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^r \quad = \quad \mathbb{E}_0 |(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^r,$$

where $\mathbb{E}_0$ denotes the expectation w.r.t. the centered measure $\mathcal{N}(0, \Sigma)$ or $\mathcal{N}(0, \Sigma_0)$.

Proof. After the first k steps $\widehat{\boldsymbol{\theta}}_k$ coincides with one of $\widetilde{\boldsymbol{\theta}}_m$, $m \le k$, and this event takes place if for some $l \le m$ the statistics $T_{l,\,m+1} > \mathfrak{z}_l$. In view of the decomposition (2.8) it holds

$$\begin{aligned}
& \{T_{l,\,m+1} > \mathfrak{z}_l \ \text{ for some } \ l = 1, \ldots, m \,|H_{m+1}\} \\
= \ & \left\{ (\widetilde{\boldsymbol{\theta}}_l - \widetilde{\boldsymbol{\theta}}_{m+1})^\top \mathbf{B}_l (\widetilde{\boldsymbol{\theta}}_l - \widetilde{\boldsymbol{\theta}}_{m+1}) > \mathfrak{z}_l \ \text{ for some } \ l = 1, \ldots, m \,|H_{m+1} \right\} \\
= \ & \left\{ \left\| \mathbf{B}_l^{1/2} \left(\mathbf{B}_l^{-1} \mathbf{\Psi} \mathbf{W}_l \Sigma_0^{1/2} \boldsymbol{\varepsilon} - \mathbf{B}_{m+1}^{-1} \mathbf{\Psi} \mathbf{W}_{m+1} \Sigma_0^{1/2} \boldsymbol{\varepsilon} \right) \right\|^2 > \mathfrak{z}_l \,, \ \ l \le m \right\}
\end{aligned}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n)$. The probability of this event does not depend on the shift $\boldsymbol{\theta}$, so without loss of generality $\boldsymbol{\theta}$ can be taken equal to zero. The risk associated with $\widehat{\boldsymbol{\theta}}_k$ admits the following decomposition:

$$\mathbb{E}_{\boldsymbol{\theta}} |(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^r = \sum_{m=1}^{k-1} \mathbb{E}_{\boldsymbol{\theta}} |(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m)|^r \mathbb{I}\{\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_m\}.$$

Under the conditions of the lemma for all $m < k$ the joint distribution of $(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m)$ does not depend on $\boldsymbol{\theta}$ by the same argumentation. □

To justify the statistical properties of the considered procedure we need the following simple observation. Let for any $\boldsymbol{\theta}$, $\boldsymbol{\theta}' \in \Theta$ the corresponding log-likelihood ratio $\mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}, \boldsymbol{\theta}')$ be defined by (2.11). Then

$$2\,\mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}, \boldsymbol{\theta}') = \|\mathbf{W}_k^{1/2}(\boldsymbol{Y} - \mathbf{\Psi}^\top \boldsymbol{\theta}')\|^2 - \|\mathbf{W}_k^{1/2}(\boldsymbol{Y} - \mathbf{\Psi}^\top \boldsymbol{\theta})\|^2.$$

Theorem 4.2. ***Quadratic shape of the fitted log-likelihood***

Let for every $k = 1, \ldots, K$ *the fitted log likelihood (FLL) be defined as follows:*

$$\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}') \stackrel{\mathrm{def}}{=} \max_{\boldsymbol{\theta} \in \Theta} \mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta}, \boldsymbol{\theta}').$$

Then

$$2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}) = (\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}). \tag{4.1}$$

Proof. Notice that $\mathrm{L}(\mathbf{W}_k, \boldsymbol{\theta})$ defined by (2.3) is quadratic in $\boldsymbol{\theta}$. The assertion follows from the Taylor expansion of the second order at the point $\widetilde{\boldsymbol{\theta}}_k$, because it is the point of maximum, and the second derivative is a constant matrix $\mathbf{B}_k$. □

Let the matrix $\mathbf{S}$ be defined as follows:

$$\mathbf{S} \stackrel{\mathrm{def}}{=} \Sigma_0^{1/2} \mathbf{W}_k \boldsymbol{\Psi}^\top \mathbf{B}_k^{-1} \boldsymbol{\Psi} \mathbf{W}_k \Sigma_0^{1/2}. \tag{4.2}$$

Then for the distribution of $\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*)$ one observes so-called "Wilks phenomenon", see [7], described by the following theorem:

Theorem 4.3. *Let the regression model be given by* (1.1) *and the parameter maximizing the expected local log-likelihood* $\boldsymbol{\theta}_k^* = \boldsymbol{\theta}_k^*(x)$ *be defined by* (2.6)*. Then for any* $k = 1, \ldots, K$ *the following equality in distribution takes place:*

$$2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*) \stackrel{\mathrm{d}}{=} \lambda_1(\mathbf{S}) \overline{\varepsilon}_1^2 + \cdots + \lambda_p(\mathbf{S}) \overline{\varepsilon}_p^2 \tag{4.3}$$

with $p = \operatorname{rank}(\mathbf{B}_k) = \dim \Theta = p$*. Here* $\lambda_1(\mathbf{S}), \ldots, \lambda_p(\mathbf{S})$ *are the non-zero eigenvalues of the matrix* $\mathbf{S}$*, and* $\overline{\varepsilon}_i$ *are independent standard normal random variables.*

Moreover, under $(A4)$ *the maximal eigenvalue* $\lambda_{max}(\mathbf{S}) \leq 1 + \delta$*, and for any* $\mathfrak{z} > 0$

$$I\!P\left\{2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*) \geq \mathfrak{z}\right\} \leq I\!P\left\{\eta \geq \mathfrak{z}/(1+\delta)\right\}, \tag{4.4}$$

where η *is a random variable distributed according to the* χ^2 *law with* p *degrees of freedom.*

Remark 4.1. Generally, if the matrix $\mathbf{B}_k$ is degenerated in (4.3) the number of terms $p \leq \dim \Theta$.

Proof. By Theorem 4.2 and the decomposition (2.8) it holds that:

$$\begin{aligned} 2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*) &= (\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) \\ &= (\mathbf{B}_k^{-1} \boldsymbol{\Psi} \mathbf{W}_k \Sigma_0^{1/2} \boldsymbol{\varepsilon})^\top \mathbf{B}_k (\mathbf{B}_k^{-1} \boldsymbol{\Psi} \mathbf{W}_k \Sigma_0^{1/2} \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\varepsilon}^\top \mathbf{S} \boldsymbol{\varepsilon}, \end{aligned}$$

where the symmetric matrix $\mathbf{S}$ is defined by (4.2). Then by the Schur theorem there exist an orthogonal matrix $\mathbf{M}$ and the diagonal matrix $\boldsymbol{\Lambda}$ composed of the eigenvalues of $\mathbf{S}$ such that $\mathbf{S} = \mathbf{M}^\top \boldsymbol{\Lambda} \mathbf{M}$. For $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n)$ and an orthogonal matrix $\mathbf{M}$ it holds that $\overline{\boldsymbol{\varepsilon}} \stackrel{\mathrm{def}}{=} \mathbf{M}\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n)$. Indeed, $\mathbb{E}\mathbf{M}\boldsymbol{\varepsilon} = \mathbb{E}\boldsymbol{\varepsilon} = 0$ and

$$\operatorname{Var} \mathbf{M}\boldsymbol{\varepsilon} = \mathbb{E}\mathbf{M}\boldsymbol{\varepsilon}(\mathbf{M}\boldsymbol{\varepsilon})^\top = \mathbf{M}\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)\mathbf{M}\boldsymbol{\varepsilon} = \mathbf{M}\mathbf{M}^\top = I_n.$$

Therefore,

$$2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*) \stackrel{\mathrm{d}}{=} \overline{\boldsymbol{\varepsilon}}^\top \boldsymbol{\Lambda} \overline{\boldsymbol{\varepsilon}}\,, \quad \overline{\boldsymbol{\varepsilon}} \sim \mathcal{N}(0, I_n)\,.$$

On the other hand, the matrix $\mathbf{S} = \Sigma_0^{1/2} \mathbf{W}_k \boldsymbol{\Psi}^\top \mathbf{B}_k^{-1} \boldsymbol{\Psi} \mathbf{W}_k \Sigma_0^{1/2}$ can be rewritten as:

$$\mathbf{S} = \Sigma_0^{1/2} \mathbf{W}_k^{1/2} \boldsymbol{\Pi}_k \mathbf{W}_k^{1/2} \Sigma_0^{1/2},$$

with $\boldsymbol{\Pi}_k = \mathbf{W}_k^{1/2} \boldsymbol{\Psi}^\top \mathbf{B}_k^{-1} \boldsymbol{\Psi} \mathbf{W}_k^{1/2}$. Notice that $\boldsymbol{\Pi}_k$ is an orthogonal projector on the linear subspace of dimension $p = \mathrm{rank}(\mathbf{B}_k)$ spanned by the rows of matrix $\boldsymbol{\Psi}$. Indeed, $\boldsymbol{\Pi}_k$ is symmetric and idempotent, i.e., $\boldsymbol{\Pi}_k^2 = \boldsymbol{\Pi}_k$.

Moreover, $\mathrm{rank}(\boldsymbol{\Pi}_k) = \mathrm{tr}(\boldsymbol{\Pi}_k) = \mathrm{tr}(\mathbf{W}_k^{1/2} \boldsymbol{\Psi}^\top \mathbf{B}_k^{-1} \boldsymbol{\Psi} \mathbf{W}_k^{1/2}) = \mathrm{tr}(\mathbf{B}_k^{-1} \boldsymbol{\Psi} \mathbf{W}_k \boldsymbol{\Psi}^\top) = \mathrm{tr}(\mathbf{B}_k^{-1} \mathbf{B}_k) = \mathrm{tr}(I_p) = p$. Therefore $\boldsymbol{\Pi}_k$ has only p unit eigenvalues and $n-p$ zero ones. Notice also that the $n \times n$ matrix $\mathbf{S}$ has $\mathrm{rank}(\mathbf{S}) = \mathrm{rank}(\boldsymbol{\Pi}_k \mathbf{W}_k^{1/2} \Sigma_0^{1/2}) = \mathrm{rank}(\boldsymbol{\Pi}_k) = p$ as well. Thus $2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*) \stackrel{\mathrm{d}}{=} \lambda_1(\mathbf{S}) \overline{\varepsilon}_1^2 + \cdots + \lambda_p(\mathbf{S}) \overline{\varepsilon}_p^2$, where $\lambda_1(\mathbf{S}), \ldots, \lambda_p(\mathbf{S})$ are the non-zero eigenvalues of the matrix $\mathbf{S}$.

Recall the definition of the matrix norm induced by the L_2 vector norm:

$$\|A\| \stackrel{\mathrm{def}}{=} \sqrt{\lambda_{max}(A^\top A)}. \tag{4.5}$$

Thus, taking into account Assumption $(A4)$, the induced L_2-norm of matrix $\mathbf{S}$ can be estimated as follows:

$$\begin{aligned}
\|\mathbf{S}\| &= \|\Sigma_0^{1/2} \mathbf{W}_k^{1/2} \boldsymbol{\Pi}_k \mathbf{W}_k^{1/2} \Sigma_0^{1/2}\| \\
&\le \|\Sigma_0^{1/2} \mathbf{W}_k^{1/2}\| \|\boldsymbol{\Pi}_k\| \|\mathbf{W}_k^{1/2} \Sigma_0^{1/2}\| \\
&= \lambda_{max}(\mathbf{W}_k \Sigma_0) \lambda_{max}(\boldsymbol{\Pi}_k) \\
&= \max_i \{ w_{k,i} \frac{\sigma_{0,i}^2}{\sigma_i^2} \} \\
&\le (1+\delta) \max_i \{ w_{k,i} \} \le 1 + \delta.
\end{aligned}$$

Therefore, the largest eigenvalue of matrix $\mathbf{S}$ is bounded: $\lambda_{max}(\mathbf{S}) \le 1 + \delta$.

The last assertion of the theorem follows from the simple observation that

$$I\!P\left\{ \lambda_1(\mathbf{S}) \overline{\varepsilon}_1^2 + \cdots + \lambda_p(\mathbf{S}) \overline{\varepsilon}_p^2 \ge \mathfrak{z} \right\} \le I\!P\left\{ \lambda_{max}(\mathbf{S}) (\overline{\varepsilon}_1^2 + \cdots + \overline{\varepsilon}_p^2) \ge \mathfrak{z} \right\}.$$

□

Corollary 4.4. ***Quasi-parametric risk bounds***

Let the model be given by (1.1) *and* $\boldsymbol{\theta}_k^* = \boldsymbol{\theta}_k^*(x)$ *be defined by* (2.6). *Assume* $(A4)$. *Then for any* $\mu < 1/(1+\delta)$

$$\mathbb{E} \exp\{ \mu\, \mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*) \} \le [1 - \mu(1+\delta)]^{-p/2} \tag{4.6}$$

$$\mathbb{E} |2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}_k^*)|^r \le (1+\delta)^r C(p, r)\,, \tag{4.7}$$

where

$$C(p, r) = \mathbb{E} |\chi_p^2|^r = 2^r \Gamma(r + p/2) / \Gamma(p/2). \tag{4.8}$$

Proof. By (4.3) and the independence of $\overline{\varepsilon}_i$

$$\begin{aligned}
\mathbb{E}\exp\{\mu\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\boldsymbol{\theta}_k^*)\} &= \mathbb{E}\exp\left\{\frac{\mu}{2}\sum_{i=1}^{p}\lambda_i(\mathbf{S})\overline{\varepsilon}_i^2\right\} \\
&= \prod_{i=1}^{p}\mathbb{E}\exp\left\{\frac{\mu}{2}\,\lambda_i(\mathbf{S})\overline{\varepsilon}_i^2\right\} \\
&= \prod_{i=1}^{p}[1-\mu\,\lambda_i(\mathbf{S})]^{-1/2} \\
&\le [1-\mu\,\lambda_{max}(\mathbf{S})]^{-p/2} \\
&\le [1-\mu(1+\delta)]^{-p/2}.
\end{aligned}$$

Let $\eta\sim\chi^2_p$. Integrating by parts yields the second inequality:

$$\begin{aligned}
\mathbb{E}|2\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\boldsymbol{\theta}_k^*)|^r &= \int_0^\infty I\!P\left\{2\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\boldsymbol{\theta}_k^*)\ge \mathfrak{z}\right\} r\mathfrak{z}^{r-1}\mathrm{d}\mathfrak{z} \\
&\le r\int_0^\infty I\!P\left\{\eta\ge \mathfrak{z}/(1+\delta)\right\}\mathfrak{z}^{r-1}\mathrm{d}\mathfrak{z} \\
&= (1+\delta)^r\,\mathbb{E}|\eta|^r.
\end{aligned}$$

□

4.2 Proof of the bounds for the critical values

Denote for any $l<k$ the variance of the difference $\widetilde{\boldsymbol{\theta}}_k-\widetilde{\boldsymbol{\theta}}_l$ by V_{lk}:

$$V_{lk}\stackrel{\mathrm{def}}{=}\mathrm{Var}(\widetilde{\boldsymbol{\theta}}_k-\widetilde{\boldsymbol{\theta}}_l)\succ 0. \tag{4.9}$$

Then there exists a unique matrix $V_{lk}^{1/2}\succ 0$ such that $(V_{lk}^{1/2})^2=V_{lk}$.

Lemma 4.5. *Assume* $(A4)$, $(A3)$ *and* $(A5)$. *If for some* $k\le K$ *the hypothesis* H_k *is true, that is, if* $\boldsymbol{\theta}_1^*=\cdots=\boldsymbol{\theta}_k^*=\boldsymbol{\theta}$, *then for any* $l<k$ *it holds that:*

$$\begin{aligned}
I\!P\left\{2\,\mathrm{L}(\mathbf{W}_l,\widetilde{\boldsymbol{\theta}}_l,\widetilde{\boldsymbol{\theta}}_k)\ge\mathfrak{z}\right\} &\le I\!P\left\{\eta\ge\mathfrak{z}/\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_lV_{lk}^{1/2})\right\} \\
&\le I\!P\left\{\eta\ge\mathfrak{z}/t_0\right\} \\
I\!P\left\{2\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\theta}}_l)\ge\mathfrak{z}\right\} &\le I\!P\left\{\eta\ge\mathfrak{z}/\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_kV_{lk}^{1/2})\right\} \\
&\le I\!P\left\{\eta\ge\mathfrak{z}/t_1\right\}
\end{aligned}$$

where $t_0=2(1+\delta)(1+u_0^{-(k-l)})$, $t_1=2(1+\delta)(1+u^{(k-l)})$, *and* η *is the* χ^2_p*-distributed random variable.*

Proof. The H_k and (2.8) imply

$$\widetilde{\boldsymbol{\theta}}_l-\widetilde{\boldsymbol{\theta}}_k=\mathbf{B}_l^{-1}\boldsymbol{\Psi}\mathbf{W}_l\Sigma_0^{1/2}\boldsymbol{\varepsilon}-\mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0^{1/2}\boldsymbol{\varepsilon}\stackrel{\mathrm{d}}{=}V_{lk}^{1/2}\xi,$$

where ξ is a standard normal vector in $I\!R^p$. Thus by Theorem 4.2 for any $l < k$

$$2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_k) = \|\mathbf{B}_l^{1/2}(\widetilde{\boldsymbol{\theta}}_l - \widetilde{\boldsymbol{\theta}}_k)\|^2 \stackrel{\mathrm{d}}{=} \xi^\top V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2}\xi.$$

By the Schur theorem there exists an orthogonal matrix M such that

$$\xi^\top V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2}\xi \stackrel{\mathrm{d}}{=} \overline{\varepsilon}^\top M^\top \Lambda_{lk} M\overline{\varepsilon},$$

where $\overline{\varepsilon}$ is a standard normal vector, $\Lambda = \mathrm{diag}(\lambda_1(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})), \cdots, \lambda_p(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2}))$, and $p = \mathrm{rank}(\mathbf{B}_l)$. Therefore,

$$2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_k) \stackrel{\mathrm{d}}{=} \lambda_1(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})\overline{\varepsilon}_1^2 + \cdots + \lambda_p(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})\overline{\varepsilon}_p^2,$$

where $\lambda_j(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})$, $j = 1, \ldots, p$, are the nonzero eigenvalues of $V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2}$.

By the similar argumentation:

$$2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_l) \stackrel{\mathrm{d}}{=} \lambda_1(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2})\overline{\varepsilon}_1^2 + \cdots + \lambda_p(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2})\overline{\varepsilon}_p^2.$$

Denote by η the χ^2_p-distributed random variable, then

$$\begin{aligned} I\!P\left\{2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_k) \geq \mathfrak{z}\right\} &\leq I\!P\left\{\eta \geq \mathfrak{z}/\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})\right\} \\ I\!P\left\{2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_l) \geq \mathfrak{z}\right\} &\leq I\!P\left\{\eta \geq \mathfrak{z}/\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2})\right\} \end{aligned}$$

For any square matrices A and B we have $(A-B)(A^\top - B^\top) \preceq 2(AA^\top + BB^\top)$. Application of this bound to the variance of the difference of estimates yields

$$\begin{aligned} V_{lk} &= \left(\mathbf{B}_l^{-1}\boldsymbol{\Psi}\mathbf{W}_l\Sigma_0^{1/2} - \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0^{1/2}\right)\left(\mathbf{B}_l^{-1}\boldsymbol{\Psi}\mathbf{W}_l\Sigma_0^{1/2} - \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0^{1/2}\right)^\top \\ &\preceq 2(\mathbf{B}_l^{-1}\boldsymbol{\Psi}\mathbf{W}_l\Sigma_0\mathbf{W}_l\boldsymbol{\Psi}^\top\mathbf{B}_l^{-1} + \mathbf{B}_k^{-1}\boldsymbol{\Psi}\mathbf{W}_k\Sigma_0\mathbf{W}_k\boldsymbol{\Psi}^\top\mathbf{B}_k^{-1}) \\ &= 2V_l + 2V_k, \end{aligned}$$

where $V_l = \mathrm{Var}\,\widetilde{\boldsymbol{\theta}}_l$, $l \leq k$. By (3.2) and by Assumption $(\mathfrak{B})$ we have:

$$\begin{aligned} V_l &\preceq (1+\delta)\mathbf{B}_l^{-1}, \\ V_k &\preceq (1+\delta)\mathbf{B}_k^{-1} \preceq (1+\delta)u_0^{-(k-l)}\mathbf{B}_l^{-1}, \\ V_{lk} &\preceq 2(1+\delta)(1+u_0^{-(k-l)})\mathbf{B}_l^{-1}. \end{aligned}$$

Therefore,

$$\mathbf{B}_l \preceq 2(1+\delta)(1+u_0^{-(k-l)})V_{lk}^{-1}. \tag{4.10}$$

Thus by (4.10) the upper bound for the induced matrix norm reads as follows:

$$\begin{aligned} \lambda_{max}(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2}) &= \|\mathbf{B}_l^{1/2}V_{lk}^{1/2}\|^2 \\ &= \sup_{\|\gamma\|=1} \gamma^\top V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2}\gamma \\ &\leq 2(1+\delta)(1+u_0^{-(k-l)}) \sup_{\|\gamma\|=1} \gamma^\top V_{lk}^{1/2}V_{lk}^{-1}V_{lk}^{1/2}\gamma \\ &\leq 2(1+\delta)(1+u_0^{-(k-l)}). \end{aligned} \tag{4.11}$$

Similarly,

$$\begin{aligned} V_{lk} &\preceq 2(1+\delta)(1+u^{(k-l)})\mathbf{B}_k^{-1}, \\ \lambda_{max}(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2}) &\leq 2(1+\delta)(1+u^{(k-l)}). \end{aligned} \tag{4.12}$$

These bounds imply

$$\begin{aligned} I\!P\left\{2\,\mathrm{L}(\mathbf{W}_l,\widetilde{\boldsymbol{\theta}}_l,\widetilde{\boldsymbol{\theta}}_k)\geq \mathfrak{z}\right\} &\leq I\!P\left\{\eta \geq \mathfrak{z}/\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})\right\} \\ &\leq I\!P\left\{\eta \geq \mathfrak{z}\left[2(1+\delta)(1+u_0^{-(k-l)})\right]^{-1}\right\} \\ I\!P\left\{2\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\theta}}_l)\geq \mathfrak{z}\right\} &\leq I\!P\left\{\eta \geq \mathfrak{z}/\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2})\right\} \\ &\leq I\!P\left\{\eta \geq \mathfrak{z}\left[2(1+\delta)(1+u^{(k-l)})\right]^{-1}\right\} \end{aligned}$$

□

Lemma 4.6. *Under the conditions of the preceding lemma for any $\mu_0 < t_0^{-1}$, or $\mu_1 < t_1^{-1}$ respectively, the exponential moments are bounded:*

$$\begin{aligned} \mathbb{E}\exp\{\mu_0\,\mathrm{L}(\mathbf{W}_l,\widetilde{\boldsymbol{\theta}}_l,\widetilde{\boldsymbol{\theta}}_k)\} &\leq [1-\mu_0 t_0]^{-p/2} \\ \mathbb{E}\exp\{\mu_1\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\theta}}_l)\} &\leq [1-\mu_1 t_1]^{-p/2}, \end{aligned}$$

where $t_0 = 2(1+\delta)(1+u_0^{-(k-l)})$ and $t_1 = 2(1+\delta)(1+u^{(k-l)})$.

Proof. The statement of the lemma is justified similarly to the proof of Corollary 4.4. The bounds (4.11) and (4.12) imply the bounds for the corresponding moment generating functions:

$$\begin{aligned} \mathbb{E}\exp\{\mu\,\mathrm{L}(\mathbf{W}_l,\widetilde{\boldsymbol{\theta}}_l,\widetilde{\boldsymbol{\theta}}_k)\} &= \prod_{j=1}^{p}\mathbb{E}\exp\{\frac{\mu}{2}\lambda_j(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})\overline{\varepsilon}_j^2\} \\ &= \prod_{j=1}^{p}[1-\mu\lambda_j(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})]^{-1/2} \\ &\leq [1-\mu\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_l V_{lk}^{1/2})]^{-p/2} \\ &\leq [1-2\mu(1+\delta)(1+u_0^{-(k-l)})]^{-p/2}\,, \\ \mathbb{E}\exp\{\mu\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\theta}}_l)\} &\leq [1-\mu\lambda_{max}(V_{lk}^{1/2}\mathbf{B}_k V_{lk}^{1/2})]^{-p/2} \\ &\leq [1-2\mu(1+\delta)(1+u^{(k-l)})]^{-p/2}. \end{aligned}$$

□

Lemma 4.7. *Under the conditions of the preceding lemma it holds that:*

$$\begin{aligned} \mathbb{E}|2\,\mathrm{L}(\mathbf{W}_l,\widetilde{\boldsymbol{\theta}}_l,\widetilde{\boldsymbol{\theta}}_k)|^r &\leq 2^r C(p,r)(1+\delta)^r(1+u_0^{-(k-l)})^r, \\ \mathbb{E}|2\,\mathrm{L}(\mathbf{W}_k,\widetilde{\boldsymbol{\theta}}_k,\widetilde{\boldsymbol{\theta}}_l)|^r &\leq 2^r C(p,r)(1+\delta)^r(1+u^{(k-l)})^r, \end{aligned}$$

where $C(p,r)$ is as in (4.8).

Proof. Integrating by parts and Lemma 4.5 yield for the second assertion

$$\begin{aligned}
\mathbb{E}|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_l)|^r &= r \int_0^\infty I\!P \left\{ 2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_l) \geq \mathfrak{z} \right\} \mathfrak{z}^{r-1} \mathrm{d}\mathfrak{z} \\
&\leq r \int_0^\infty I\!P \left\{ \eta \geq \mathfrak{z} \left[2(1+\delta)(1+u^{(k-l)}) \right]^{-1} \right\} \mathfrak{z}^{r-1} \mathrm{d}\mathfrak{z} \\
&= 2^r (1+\delta)^r (1+u^{(k-l)})^r I\!E|\eta|^r,
\end{aligned}$$

where $\eta \sim \chi^2_p$. The first assertion is proved similarly. □

Proof. of Theorem 3.1 The theoretical choice of the critical values The risk corresponding to the adaptive estimate can be represented as a sum of risks of the false alarms at each step of the procedure:

$$\mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^r = \sum_{m=1}^{k-1} \mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m)|^r \mathbb{I}\{\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_m\}.$$

By the definition of the last accepted estimate $\widehat{\boldsymbol{\theta}}_k$, for any $m = 1, \ldots, k-1$, the event $\{\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_m\}$ happens if for some $l = 1, \ldots, m$ the statistic $T_{l,m+1} > \mathfrak{z}_l$. Thus

$$\{\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_m\} \subseteq \bigcup_{l=1}^{m} \{T_{l,m+1} > \mathfrak{z}_l\}.$$

It holds also that for any positive μ

$$\begin{aligned}
\mathbb{I}\{T_{l,m+1} > \mathfrak{z}_l\} &= \mathbb{I}\{2\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_{m+1}) - \mathfrak{z}_l > 0\} \\
&\leq \exp\{\frac{\mu}{2}\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_{m+1}) - \frac{\mu}{4}\mathfrak{z}_l\}.
\end{aligned}$$

This simple fact and the Cauchy-Schwarz inequality imply for $m = 1, \ldots, k-1$ the following bound:

$$\begin{aligned}
& \mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widetilde{\boldsymbol{\theta}}_m)|^r \mathbb{I}\{\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_m\} \\
= \;& \mathbb{E}_{0,\Sigma}|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_m)|^r \mathbb{I}\{\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_m\} \\
\leq \;& \sum_{l=1}^{m} e^{-\frac{\mu}{4}\mathfrak{z}_l} \mathbb{E}_{0,\Sigma} \left[|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_m)|^r \exp\{\frac{\mu}{2}\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_{m+1})\} \right] \\
\leq \;& \sum_{l=1}^{m} e^{-\frac{\mu}{4}\mathfrak{z}_l} \left\{ \mathbb{E}_{0,\Sigma} \left[|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_m)|^{2r} \right] \right\}^{1/2} \left\{ \mathbb{E}_{0,\Sigma} \left[\exp\{\mu\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_{m+1})\} \right] \right\}^{1/2}.
\end{aligned}$$

By Lemma 4.6 with $\delta = 0$

$$\mathbb{E}_{0,\Sigma} \left[\exp\{\mu\,\mathrm{L}(\mathbf{W}_l, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_{m+1})\} \right] < (1-4\mu)^{-p/2}.$$

This together with the bound from Lemma 4.7 gives

$$\begin{aligned}
& \mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)^\top \mathbf{B}_k (\widetilde{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_k)|^r \\
\leq & \ 2^r \sqrt{C(p,2r)}(1-4\mu)^{-p/4} \sum_{m=1}^{k-1} \sum_{l=1}^{m} e^{-\frac{\mu}{4}\mathfrak{z}_l} (1 + u^{(k-m)})^r \\
= & \ 2^r \sqrt{C(p,2r)}(1-4\mu)^{-p/4} \sum_{l=1}^{k-1} e^{-\frac{\mu}{4}\mathfrak{z}_l} \sum_{m=l}^{k-1} (1 + u^{(k-m)})^r \\
\leq & \ 2^{2r} \sqrt{C(p,2r)}(1-4\mu)^{-p/4} (1-u^{-r})^{-1} \sum_{l=1}^{k-1} e^{-\frac{\mu}{4}\mathfrak{z}_l} u^{r(k-l)},
\end{aligned}$$

because $-(k-l) < -(m-l)$ and

$$\begin{aligned}
\sum_{m=l}^{k-1} (1 + u^{(k-m)})^r &= u^{r(k-l)} \sum_{m=l}^{k-1} (u^{-(k-l)} + u^{-(m-l)})^r \\
&< 2^r u^{r(k-l)} \sum_{m=l}^{k-1} u^{-r(m-l)} \\
&< 2^r u^{r(k-l)} (1-u^{-r})^{-1}.
\end{aligned}$$

Since $u^{r(k-l)} \leq u^{r(K-l)}$ for any $l < k \leq K$ the choice

$$\mathfrak{z}_l = \frac{4}{\mu} \Big\{ r(K-l)\log u + \log(K/\alpha) - \frac{p}{4}\log(1-4\mu) - \log(1-u^{-r}) + \overline{C}(p,r) \Big\}.$$

with

$$\overline{C}(p,r) = \log \left\{ \frac{2^{2r}[\Gamma(2r+p/2)\Gamma(p/2)]^{1/2}}{\Gamma(r+p/2)} \right\}$$

provides the required bound

$$\mathbb{E}_{0,\Sigma}|(\widetilde{\boldsymbol{\theta}}_l - \widehat{\boldsymbol{\theta}}_l)^\top \mathbf{B}_l (\widetilde{\boldsymbol{\theta}}_l - \widehat{\boldsymbol{\theta}}_l)|^r \leq \alpha C(p,r) \quad \text{for all } l = 2, \ldots, K.$$

□

4.3 Matrix results

Lemma 4.8. *The matrices $J_k \otimes \Sigma$ and $J_k \otimes \Sigma_0$ are positive semidefinite for any $k = 2, \ldots, K$.*

Moreover, under Assumption $(A4)$ with the same δ, the similar to $(A4)$ relation holds for the covariance matrices $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Sigma}_{k,0}$ of linear estimates:

$$(1-\delta)\boldsymbol{\Sigma}_k \preceq \boldsymbol{\Sigma}_{k,0} \preceq (1+\delta)\boldsymbol{\Sigma}_k, \quad k \leq K.$$

Proof. Symmetry of J_k and Σ, (respectively, Σ_0) implies symmetry of $J_k \otimes \Sigma$, (respectively, $J_k \otimes \Sigma_0$). Notice that any vector $\gamma_{nk} \in I\!R^{nk}$ can be represented as a partitioned vector $\gamma_{nk}^\top = ((\gamma_{nk}^{(1)})^\top, (\gamma_{nk}^{(2)})^\top, \ldots, (\gamma_{nk}^{(k)})^\top)$, with $\gamma_{nk}^{(l)} \in I\!R^n$, $l = 1, \ldots, k$. Then

$$\gamma_{nk}^\top (J_k \otimes \Sigma)\gamma_{nk} = \big(\sum_{l=1}^{k} \gamma_{nk}^{(l)}\big)^\top \Sigma \big(\sum_{l=1}^{k} \gamma_{nk}^{(l)}\big) = \widetilde{\gamma}_n^\top \, \Sigma \, \widetilde{\gamma}_n, \tag{4.13}$$

where $\widetilde{\gamma}_n \stackrel{\mathrm{def}}{=} \sum_{l=1}^{k} \gamma_{nk}^{(l)} \in I\!R^n$. Because $\Sigma \succ 0$ it implies $\widetilde{\gamma}_n^\top \Sigma \widetilde{\gamma}_n > 0$ for all $\widetilde{\gamma}_n \neq 0$. But even for $\gamma_{nk} \neq 0$, if its subvectors $\{\gamma_{nl}^{(l)}\}$ are linearly dependent, $\widetilde{\gamma}_n$ can be zero. Thus there exists a nonzero vector γ such that $\gamma^\top (J_k \otimes \Sigma)\gamma = 0$. This means positive semidefiniteness.

The second assertion follows from the observation that Assumption $(A4)$ due to the equality (4.13) also holds for the Kronecker product

$$(1-\delta) J_k \otimes \Sigma \preceq J_k \otimes \Sigma_0 \preceq (1+\delta) J_k \otimes \Sigma. \tag{4.14}$$

Therefore

$$(1-\delta)\mathbf{D}_k (J_k \otimes \Sigma)\mathbf{D}_k^\top \preceq \mathbf{D}_k (J_k \otimes \Sigma_0)\mathbf{D}_k^\top \preceq (1+\delta)\mathbf{D}_k (J_k \otimes \Sigma)\mathbf{D}_k^\top .$$

□

Lemma 4.9. *Fix $x \in I\!R^d$. Suppose that the weights $\{w_{l,\,i}(x)\}$ satisfy*

$$w_{l,\,i}(x) w_{m,\,i}(x) = w_{l,\,i}(x)\ , \quad l \leq m. \tag{4.15}$$

Then under Assumptions $(A1)$, $(A2)$, $(A5)$ the covariance matrix $\mathbf{\Sigma}_k$ defined by (3.5) *is nonsingular with*

$$\det \mathbf{\Sigma}_k = \det \mathbf{B}_k^{-1} \prod_{l=2}^{k} \det(\mathbf{B}_{l-1}^{-1} - \mathbf{B}_l^{-1}) > 0\ , \quad k = 2, \ldots, K. \tag{4.16}$$

Remark 4.2. The condition (4.15) holds for rectangular kernels with nested supports.

Proof. The condition (4.15) implies

$$\mathbf{W}_l \Sigma \mathbf{W}_m = \mathrm{diag}(w_{l,1} w_{m,1}/\sigma_1^2, \ldots, w_{l,n} w_{m,n}/\sigma_n^2) = \mathbf{W}_l$$

for any $l \leq m$. Thus the blocks of $\mathbf{\Sigma}_k$ simplify to

$$D_l \Sigma D_m^\top = \mathbf{B}_l^{-1} \mathbf{\Psi} \mathbf{W}_l \Sigma \mathbf{W}_m \mathbf{\Psi}^\top \mathbf{B}_m^{-1} = \mathbf{B}_l^{-1} \mathbf{\Psi} \mathbf{W}_l \mathbf{\Psi}^\top \mathbf{B}_m^{-1}$$

and $\mathbf{\Sigma}_k$ has a simple structure:

$$\mathbf{\Sigma}_k = \begin{pmatrix} \mathbf{B}_1^{-1} & \mathbf{B}_2^{-1} & \mathbf{B}_3^{-1} & \ldots & \mathbf{B}_k^{-1} \\ \mathbf{B}_2^{-1} & \mathbf{B}_2^{-1} & \mathbf{B}_3^{-1} & \ldots & \mathbf{B}_k^{-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{B}_k^{-1} & \mathbf{B}_k^{-1} & \mathbf{B}_k^{-1} & \ldots & \mathbf{B}_k^{-1} \end{pmatrix}.$$

Then the determinant of $\mathbf{\Sigma}_k$ coincides with the determinant of the following irreducible block triangular matrix:

$$\det \mathbf{\Sigma}_k = \begin{vmatrix} \mathbf{B}_1^{-1} - \mathbf{B}_2^{-1} & \mathbf{B}_2^{-1} - \mathbf{B}_3^{-1} & \ldots & \mathbf{B}_{k-1}^{-1} - \mathbf{B}_k^{-1} & \mathbf{B}_k^{-1} \\ \mathbf{0} & \mathbf{B}_2^{-1} - \mathbf{B}_3^{-1} & \ldots & \mathbf{B}_{k-1}^{-1} - \mathbf{B}_k^{-1} & \mathbf{B}_k^{-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{B}_{k-1}^{-1} - \mathbf{B}_k^{-1} & \mathbf{B}_k^{-1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_k^{-1} \end{vmatrix}$$

implying

$$\det \boldsymbol{\Sigma}_k = \det(\mathbf{B}_1^{-1} - \mathbf{B}_2^{-1}) \det(\mathbf{B}_2^{-1} - \mathbf{B}_3^{-1}) \cdot \ldots \cdot \det(\mathbf{B}_{k-1}^{-1} - \mathbf{B}_k^{-1}) \det \mathbf{B}_k^{-1}.$$

Clearly the matrix $\boldsymbol{\Sigma}_k$ is nonsingular if all the matrices $\mathbf{B}_{l-1}^{-1} - \mathbf{B}_l^{-1}$ are nonsingular. By $(A1)$ and $(A2)$ $\mathbf{B}_l \succ 0$ for any l. By $(A5)$ there exists $u_0 > 1$ such that $\mathbf{B}_l \succeq u_0 \mathbf{B}_{l-1}$, therefore $\mathbf{B}_{l-1}^{-1} - \mathbf{B}_l^{-1} \succeq (1 - 1/u_0)\mathbf{B}_{l-1}^{-1} \succ \mathbf{B}_{l-1}^{-1} \succ 0$. □

Lemma 4.10. *In the "nonparametric situation" the moment generation function (mgf) of the joint distribution of* $\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_K$ *is*

$$\mathbb{E} \exp\bigl\{\gamma^\top (\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_K - \operatorname{vec} \boldsymbol{\Theta}_K^*)\bigr\} = \exp\left\{\frac{1}{2}\gamma^\top \boldsymbol{\Sigma}_{K,0}\, \gamma\right\}. \tag{4.17}$$

Thus, provided that $\boldsymbol{\Sigma}_{K,0} \succ 0$, *it holds that* $\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_K \sim \mathcal{N}(\operatorname{vec} \boldsymbol{\Theta}_K^*, \boldsymbol{\Sigma}_{K,0})$.

Similarly, in the "parametric situation", if $\boldsymbol{\Sigma}_K \succ 0$, *then the joint distribution of* $\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_K$ *is* $\mathcal{N}(\operatorname{vec} \boldsymbol{\Theta}_K, \boldsymbol{\Sigma}_K)$ *with the mgf:*

$$\mathbb{E} \exp\bigl\{\gamma^\top (\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_K - \operatorname{vec} \boldsymbol{\Theta}_K)\bigr\} = \exp\left\{\frac{1}{2}\gamma^\top \boldsymbol{\Sigma}_K\, \gamma\right\}. \tag{4.18}$$

Proof. Let $\gamma \in \mathbb{R}^{pK}$ be written in a partitioned form $\gamma^\top = (\gamma_1^\top, \ldots, \gamma_K^\top)$ with $\gamma_l \in \mathbb{R}^p$, $l = 1, \ldots, K$. Then the mgf for the centered random vector $\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_K - \operatorname{vec} \boldsymbol{\Theta}_K^* \in \mathbb{R}^{pK}$, due to the decomposition (2.8) $\widetilde{\boldsymbol{\theta}}_l = \boldsymbol{\theta}_l^* + D_l \Sigma_0^{1/2} \boldsymbol{\varepsilon}$ with $D_l = \mathbf{B}_l^{-1} \boldsymbol{\Psi} \mathbf{W}_l$, can be represented as follows:

$$\begin{aligned} & \mathbb{E} \exp\bigl\{\gamma^\top (\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_K - \operatorname{vec} \boldsymbol{\Theta}_K^*)\bigr\} = \mathbb{E} \exp\bigl\{\sum_{l=1}^K \gamma_l^\top (\widetilde{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l^*)\bigr\} \\ = \quad & \mathbb{E} \exp\bigl\{\sum_{l=1}^K \gamma_l^\top D_l \Sigma_0^{1/2} \boldsymbol{\varepsilon}\bigr\} = \mathbb{E} \exp\bigl\{\bigl(\sum_{l=1}^K D_l^\top \gamma_l\bigr)^\top \Sigma_0^{1/2} \boldsymbol{\varepsilon}\bigr\}. \end{aligned}$$

A trivial observation that $\sum_{l=1}^K D_l^\top \gamma_l$ is a vector in $\mathbb{R}^n$ and $\Sigma_0^{1/2} \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma_0)$ by (1.1) implies by the definition of $\boldsymbol{\Sigma}_{K,0}$ the first assertion of the lemma, because

$$\begin{aligned} & \mathbb{E} \exp\bigl\{\bigl(\sum_{l=1}^K D_l^\top \gamma_l\bigr)^\top \Sigma_0^{1/2} \boldsymbol{\varepsilon}\bigr\} = \exp\left\{\frac{1}{2}\bigl(\sum_{l=1}^K D_l^\top \gamma_l\bigr)^\top \Sigma_0 \bigl(\sum_{l=1}^K D_l^\top \gamma_l\bigr)\right\} \\ = \quad & \exp\left\{\frac{1}{2}\bigl(\mathbf{D}_K^\top \gamma\bigr)^\top (J_K \otimes \Sigma_0) \mathbf{D}_K^\top \gamma\right\} = \exp\left\{\frac{1}{2}\gamma^\top \boldsymbol{\Sigma}_{K,0}\, \gamma\right\}, \end{aligned}$$

here $\mathbf{D}_K$ is defined by (3.23). □

4.4 Proof of the propagation property

Lemma 4.11. *The Kullback-Leibler divergence between the distributions of* $\operatorname{vec} \widetilde{\boldsymbol{\Theta}}_k$ *under the alternative and under the null has the following form:*

$$\begin{aligned} & 2\mathbb{KL}(I\!P_{\boldsymbol{f},\Sigma_0}^k, I\!P_{\boldsymbol{\theta},\Sigma}^k) \stackrel{\mathrm{def}}{=} 2\mathbb{E}_{\boldsymbol{f},\Sigma_0} \log\bigl(\frac{\mathrm{d}I\!P_{\boldsymbol{f},\Sigma_0}^k}{\mathrm{d}I\!P_{\boldsymbol{\theta},\Sigma}^k}\bigr) \\ = \quad & \Delta(k) + \log\left(\frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}}\right) + \operatorname{tr}(\boldsymbol{\Sigma}_k^{-1} \boldsymbol{\Sigma}_{k,0}) - pk, \end{aligned} \tag{4.19}$$

where

$$b(k) \stackrel{\text{def}}{=} \operatorname{vec}\boldsymbol{\Theta}_k^* - \operatorname{vec}\boldsymbol{\Theta}_k \tag{4.20}$$

$$\Delta(k) \stackrel{\text{def}}{=} b(k)^\top \boldsymbol{\Sigma}_k^{-1} b(k). \tag{4.21}$$

Proof. Denote the Radon-Nikodym derivative by $Z_k \stackrel{\text{def}}{=} \mathrm{d}I\!P^k_{\boldsymbol{f},\Sigma_0}/\mathrm{d}I\!P^k_{\boldsymbol{\theta},\Sigma}$. Then

$$\begin{aligned}\log\bigl(Z_k(y)\bigr) = \frac{1}{2}\log\left(\frac{\det\boldsymbol{\Sigma}_k}{\det\boldsymbol{\Sigma}_{k,0}}\right) &- \frac{1}{2}\|\boldsymbol{\Sigma}_{k,0}^{-1/2}(y - \operatorname{vec}\boldsymbol{\Theta}_k^*)\|^2 \\ &+ \frac{1}{2}\|\boldsymbol{\Sigma}_k^{-1/2}(y - \operatorname{vec}\boldsymbol{\Theta}_k)\|^2 \end{aligned} \tag{4.22}$$

can be considered as a quadratic function of $\operatorname{vec}\boldsymbol{\Theta}_k$. By the Taylor expansion at the point $\operatorname{vec}\boldsymbol{\Theta}_k^*$ the last expression reads as follows

$$\begin{aligned}\log\bigl(Z_k(y)\bigr) &= \frac{1}{2}\log\left(\frac{\det\boldsymbol{\Sigma}_k}{\det\boldsymbol{\Sigma}_{k,0}}\right) - \frac{1}{2}\|\boldsymbol{\Sigma}_{k,0}^{-1/2}(y - \operatorname{vec}\boldsymbol{\Theta}_k^*)\|^2 \\ &+ \frac{1}{2}\|\boldsymbol{\Sigma}_k^{-1/2}(y - \operatorname{vec}\boldsymbol{\Theta}_k^*)\|^2 + b(k)^\top\boldsymbol{\Sigma}_k^{-1}(y - \operatorname{vec}\boldsymbol{\Theta}_k^*) + \frac{1}{2}\Delta(k).\end{aligned}$$

Then the expression for the Kullback-Leibler divergence can be written in the following way:

$$\begin{aligned}&\mathbb{KL}(I\!P^k_{\boldsymbol{f},\Sigma_0}, I\!P^k_{\boldsymbol{\theta},\Sigma}) \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{f},\Sigma_0}\log\bigl(Z_k\bigr) \\ = \;& \frac{1}{2}\log\left(\frac{\det\boldsymbol{\Sigma}_k}{\det\boldsymbol{\Sigma}_{k,0}}\right) + \frac{1}{2}\Delta(k) + \frac{1}{2}\mathbb{E}\bigl\{\|\boldsymbol{\Sigma}_k^{-1/2}\boldsymbol{\Sigma}_{k,0}^{1/2}\xi\|^2 - \|\xi\|^2 + 2b(k)^\top\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{k,0}^{1/2}\xi\bigr\},\end{aligned}$$

where $\xi \sim \mathcal{N}(0, I_{pk})$. This implies

$$2\mathbb{KL}(I\!P^k_{\boldsymbol{f},\Sigma_0}, I\!P^k_{\boldsymbol{\theta},\Sigma}) = \Delta(k) + \log\left(\frac{\det\boldsymbol{\Sigma}_k}{\det\boldsymbol{\Sigma}_{k,0}}\right) + \operatorname{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{k,0}) - pk. \tag{4.23}$$

In the case of *homogeneous errors* with $\sigma_{0,i} = \sigma_0$ and $\sigma_i = \sigma, i = 1, \ldots, n$ the calculations simplify a lot. Now

$$\boldsymbol{\Sigma}_k = \sigma^2\mathbf{V}_k, \quad \boldsymbol{\Sigma}_{k,0} = \sigma_0^2\mathbf{V}_k$$

with a $pk \times pk$ matrix $\mathbf{V}_k$ defined as

$$\mathbf{V}_k = \bigl(\overline{D}_1 \oplus \cdots \oplus \overline{D}_k\bigr)\bigl(J_k \otimes I_n\bigr)\bigl(\overline{D}_1 \oplus \cdots \oplus \overline{D}_k\bigr)^\top,$$

where $\overline{D}_l = (\boldsymbol{\Psi}\mathcal{W}_l\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}\mathcal{W}_l$, $l = 1, \ldots, k$ does not depend on σ. Then $\Delta(k) = \sigma^{-2}\Delta_1(k)$, with $\Delta_1(k) \stackrel{\text{def}}{=} b(k)^\top\mathbf{V}_k^{-1}b(k)$, $\det\boldsymbol{\Sigma}_k/\det\boldsymbol{\Sigma}_{k,0} = (\sigma^2/\sigma_0^2)^{pk}$, and the expression for the Kullback-Leibler divergence reads as follows:

$$\begin{aligned}\mathbb{KL}(I\!P^k_{\boldsymbol{f},\Sigma_0}, I\!P^k_{\boldsymbol{\theta},\Sigma}) &= pk\log\bigl(\frac{\sigma}{\sigma_0}\bigr) + \frac{1}{2}\Delta(k) + \frac{pk}{2}\bigl(\frac{\sigma_0^2}{\sigma^2} - 1\bigr) \\ &= pk\log\bigl(\frac{\sigma}{\sigma_0}\bigr) + \frac{1}{2\sigma^2}b(k)^\top\mathbf{V}_k^{-1}b(k) + \frac{pk}{2}\bigl(\frac{\sigma_0^2}{\sigma^2} - 1\bigr),\end{aligned} \tag{4.24}$$

implying the same asymptotic behavior as in (3.13).

□

Proof. of Theorem 3.2 (Propagation property)

Notice that for any nonnegative measurable function $g = g(\widetilde{\boldsymbol{\Theta}}_k)$ the Cauchy-Schwarz inequality implies

$$\mathbb{E}_{\boldsymbol{f},\Sigma_0}[g] = \mathbb{E}_{\boldsymbol{\theta},\Sigma}[gZ_k] \le \big(\mathbb{E}_{\boldsymbol{\theta},\Sigma}[g^2]\big)^{1/2}\big(\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2]\big)^{1/2} \tag{4.25}$$

with the Radon-Nikodym derivative $Z_k = \mathrm{d}I\!P^k_{\boldsymbol{f},\Sigma_0}/\mathrm{d}I\!P^k_{\boldsymbol{\theta},\Sigma}$. One gets the first assertion taking $g = |(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})^\top \mathbf{B}_k(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})|^{r/2}$, and applying "the parametric risk bound" with $\delta = 0$ from (4.7):

$$\begin{aligned}
\mathbb{E}[g] &\le \big(\mathbb{E}_{\boldsymbol{\theta},\Sigma}|(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})^\top \mathbf{B}_k(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta})|^r\big)^{1/2}\big(\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2]\big)^{1/2} \\
&= \big(\mathbb{E}_{\boldsymbol{\theta},\Sigma}|2\,\mathrm{L}(\mathbf{W}_k, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta})|^r\big)^{1/2}\big(\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2]\big)^{1/2} \\
&\le (\mathbb{E}|\chi^2_p|^r)^{1/2}\big(\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2]\big)^{1/2}.
\end{aligned}$$

The second assertion is treated similarly by application of the pivotality property from Lemma 4.1 and the propagation conditions (2.17).

To calculate $\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2]$ let us consider $\log Z_k$ given by

$$\begin{aligned}
\log\big(Z_k(y)\big) = \frac{1}{2}\log\left(\frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}}\right) &- \frac{1}{2}\|\boldsymbol{\Sigma}_{k,0}^{-1/2}(y - \operatorname{vec}\boldsymbol{\Theta}^*_k)\|^2 \\
&+ \frac{1}{2}\|\boldsymbol{\Sigma}_{k}^{-1/2}(y - \operatorname{vec}\boldsymbol{\Theta}_k)\|^2
\end{aligned}$$

as a function of $\operatorname{vec}\boldsymbol{\Theta}^*_k$. Application of the Taylor expansion at the point $\operatorname{vec}\boldsymbol{\Theta}_k$ yields

$$\begin{aligned}
2\log Z_k &= \log\frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}} - \|\boldsymbol{\Sigma}_{k,0}^{-1/2}(y - \operatorname{vec}\boldsymbol{\Theta}_k)\|^2 + \|\boldsymbol{\Sigma}_{k}^{-1/2}(y - \operatorname{vec}\boldsymbol{\Theta}_k)\|^2 \\
&+ 2b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1}(y - \operatorname{vec}\boldsymbol{\Theta}_k) - b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1} b(k).
\end{aligned}$$

With $\xi \sim \mathcal{N}(0, I_{pk})$ the second moment of the Radon-Nikodym derivative reads as follows

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2] \\
&= \frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}}\exp\{-b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1} b(k)\}\mathbb{E}\exp\{-\|\boldsymbol{\Sigma}_{k,0}^{-1/2}\boldsymbol{\Sigma}_k^{1/2}\xi\|^2 + \|\xi\|^2 + 2b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2}\xi\} \\
&= \frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}}\big[\det\big(2\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2} - I_{pk}\big)\big]^{-1/2} \\
&\times \exp\{2b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2}\big(2\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2} - I_{pk}\big)^{-1}\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1} b(k) - b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1} b(k)\} \\
&= \frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}}\big[\prod_{j=1}^{pk}\{2\lambda_j(\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2}) - 1\}\big]^{-1/2} \qquad (4.26) \\
&\times \exp\{b(k)^\top \boldsymbol{\Sigma}_{k,0}^{-1/2}\big[2\boldsymbol{\Sigma}_{k,0}^{-1/2}\boldsymbol{\Sigma}_k^{1/2}\big(2\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2} - I_{pk}\big)^{-1}\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1/2} - I_{pk}\big]\boldsymbol{\Sigma}_{k,0}^{-1/2} b(k)\}.
\end{aligned}$$

To estimate the obtained expression in terms of the level of noise misspecification δ notice that the condition (3.8) implies

$$\left(\frac{1}{1+\delta}\right)^{pk} \le \frac{\det \boldsymbol{\Sigma}_k}{\det \boldsymbol{\Sigma}_{k,0}} \le \left(\frac{1}{1-\delta}\right)^{pk},$$

$$\left(\frac{1-\delta}{1+\delta}\right)^{\frac{pk}{2}} \le \big[\prod_{j=1}^{pk}\{2\lambda_j(\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2})-1\}\big]^{-1/2} \le \left(\frac{1+\delta}{1-\delta}\right)^{\frac{pk}{2}}.$$

$$\frac{1-\delta}{1+\delta}I_{pk} \preceq \left(2\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2}-I_{pk}\right)^{-1} \preceq \frac{1+\delta}{1-\delta}I_{pk}.$$

Therefore the quantity in the exponent in (4.26) is bounded by:

$$\begin{aligned}
&\left(2\frac{1-\delta}{(1+\delta)^2}-1\right)b(k)^\top\boldsymbol{\Sigma}_{k,0}^{-1}b(k)\\
&\le b(k)^\top\boldsymbol{\Sigma}_{k,0}^{-1/2}\big[2\boldsymbol{\Sigma}_{k,0}^{-1/2}\boldsymbol{\Sigma}_k^{1/2}\big(2\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1}\boldsymbol{\Sigma}_k^{1/2}-I_{pk}\big)^{-1}\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Sigma}_{k,0}^{-1/2}-I_{pk}\big]\boldsymbol{\Sigma}_{k,0}^{-1/2}b(k)\\
&\le \left(2\frac{1+\delta}{(1-\delta)^2}-1\right)b(k)^\top\boldsymbol{\Sigma}_{k,0}^{-1}b(k).
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\frac{\Delta(k)}{1+\delta} = \frac{1}{1+\delta}b(k)^\top\boldsymbol{\Sigma}_k^{-1}b(k)\\
&\le b(k)^\top\boldsymbol{\Sigma}_{k,0}^{-1}b(k)\\
&\le \frac{1}{1-\delta}b(k)^\top\boldsymbol{\Sigma}_k^{-1}b(k) = \frac{\Delta(k)}{1-\delta}.
\end{aligned}$$

Finally,

$$\begin{aligned}
&\left(\frac{1-\delta}{(1+\delta)^3}\right)^{\frac{pk}{2}}\exp\left\{\left(\frac{2(1-\delta)}{(1+\delta)^2}-1\right)\frac{\Delta(k)}{1+\delta}\right\}\\
&\le \mathbb{E}_{\boldsymbol{\theta},\Sigma}[Z_k^2] \le \left(\frac{1+\delta}{(1-\delta)^3}\right)^{\frac{pk}{2}}\exp\left\{\left(\frac{2(1+\delta)}{(1-\delta)^2}-1\right)\frac{\Delta(k)}{1-\delta}\right\}. \qquad (4.27)
\end{aligned}$$

In the *case of homogeneous errors* the expression for $\log Z_k$ reads as

$$\begin{aligned}
\log Z_k &= pk\log\big(\frac{\sigma}{\sigma_0}\big)+\frac{1}{2}\big(\frac{1}{\sigma^2}-\frac{1}{\sigma_0^2}\big)\|\mathbf{V}_k^{-1/2}(y-\operatorname{vec}\boldsymbol{\Theta}_k)\|^2\\
&+ \frac{1}{\sigma_0^2}b(k)^\top\mathbf{V}_k^{-1}(y-\operatorname{vec}\boldsymbol{\Theta}_k)-\frac{1}{2\sigma_0^2}b(k)^\top\mathbf{V}_k^{-1}b(k),
\end{aligned}$$

implying

$$\mathbb{E}_{\boldsymbol{\theta},\sigma}[Z_k^2] = \left(\frac{\sigma^2}{\sigma_0^2}\right)^{pk}\left(\frac{\sigma_0^2}{2\sigma^2-\sigma_0^2}\right)^{\frac{pk}{2}}\exp\left\{\frac{b(k)^\top\mathbf{V}_k^{-1}b(k)}{2\sigma^2-\sigma_0^2}\right\}.$$

By Assumption $(A4)$

$$\begin{aligned}
&\left(\frac{1-\delta}{(1+\delta)^3}\right)^{\frac{pk}{2}}\exp\left\{\frac{\Delta_1(k)}{\sigma^2(1+\delta)}\right\}\\
\le \mathbb{E}_{\boldsymbol{\theta},\sigma}[Z_k^2] &\le \left(\frac{1+\delta}{(1-\delta)^3}\right)^{\frac{pk}{2}}\exp\left\{\frac{\Delta_1(k)}{\sigma^2(1-\delta)}\right\}, \qquad (4.28)
\end{aligned}$$

where p is the dimension of the parameter set and k is the degree of the localization. □

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory (Tsahkadsor, 1971), Akadmiai Kiad, Budapest* 267–281.

[2] Baraud, Y., Giraud,C. and Huet,S. (2009) Gaussian model selection with an unknown variance. *Ann. Statist.* **37:2** 630–672.

[3] Birgé, L. and Massart, P. (2001) Gaussian model selection. *Journal of the European Mathematical Society* **3:3** 203–268.

[4] Brua, J.-Y. (2009). Asymptotic efficient estimators for non-parametric heteroscedastic model. *Statistical Methodology* **6:1** 47–60.

[5] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrica* **81** 425–455

[6] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Monographs on Statistics and Applied Probability, 66. Chapman and Hall, London.

[7] Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29:1** 153–193.

[8] Galtchouk, L. and Pergamenshchikov, S. (2009). Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression *Journal of the Korean Statistical Society* **38:4** 305–322.

[9] Galtchouk, L. and Pergamenshchikov, S. (2010). Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression. arXiv:1002.1537v1

[10] Galtchouk, L. and Pergamenshchikov, S. (2010). Sharp non-asymptotic oracle inequalities for nonparametric heteroscedastic regression models. arXiv:1002.1538v1

[11] Goldenshluger, A. and Nemirovski, A. (1994). On spatial adaptive estimation of nonparametric regression. *Research report, Technion-Israel Inst. Technology, Haifa, Israel*.

[12] Efroimovich, S. and Pinsker, M.(1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statistica Sinica* **6** 925–942.

[13] Efroimovich, S. (2007). Sequential design and estimation in het- eroscedastic nonparametric regression. *Sequential Analysis* **26** 3–25.

[14] Katkovnik, V. Ja. (1979). Linear and nonlinear methods of nonparametric regression analysis. (Russian) *Soviet Automat. Control* **5** 35–46, 93.

[15] Katkovnik, V. Ja. (1983). Convergence of linear and nonlinear nonparametric estimates of "kernel" type. *Automat. Remote Control* **44:4** 495–506; translated from *Avtomat. i Telemekh.* 1983 **4** 108–120 (Russian).

[16] Katkovnik, V. Ja. (1985). *Nonparametric Identification and Data Smoothing: Local Approximation Approach*. Nauka, Moscow (Russian).

[17] Katkovnik, V., Egiazarian, K. and Astola, J. (2006). *Local Approximation Techniques in Signal and Image Processing*. Bellingham, WA: SPIE Press.

[18] Katkovnik, V. and Spokoiny, V. (2008). Spatially adaptive estimation via fitted local likelihood techniques. *IEEE Trans. Signal Process.*, **56**, No.3, 873–886.

[19] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics* **22** 79–86.

[20] Lepskii, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. (Russian) *Teor. Veroyatnost. i Primenen.* **35:3** 459–470; translation in *Theory Probab. Appl.* **35:3** 454–466.

[21] Lepskii, O. V. (1992). Asymptotic minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. (Russian) *Teor. Veroyatnost. i Primenen.* **37:3** 468–481; translation in *Theory Probab. Appl.* **37:3** 433–448.

[22] Lepski, O. V., Mammen, E. and Spokoiny, V.G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Stat.* **25:3** 929–947.

[23] Lepski, O. V. and Spokoiny, V.G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Stat.* **25:6** 2512–2546.

[24] Loader, C. (1999). *Local Regression and Likelihood.* Statistics and Computing. Springer-Verlag, New York.

[25] Massart, P. (2003) *Concentration Inequalities and Model Selection* (2007). Ecole d'été de Probabilités de Saint-Flour . Lecture Notes in Mathematics 1896, Springer Berlin/Heidelberg.

[26] Spokoiny, V. (2002). Variance estimation for high-dimensional regression models. *J. Multivariate Anal.* **82** 111–133.

[27] Spokoiny, V. and Vial, C. (2009). Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.* **37:5B** 2783–2807.

[28] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer Series in Statistics. Springer-Verlag, New York.

[29] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50:1** 1–25.